УДК 681.3.07

НЕКОТОРЫЕ АСПЕКТЫ СТАТИСТИЧЕСКОГО АНАЛИЗА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМИЧЕСКОГО ЯЗЫКА R

Смагин Борис Игнатьевич
доктор экономических наук, профессор
bismagin@mail.ru
Мичуринский государственный аграрный университет
Мичуринск, Россия

Аннотация. R – язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда с открытым исходным кодом, позволяет формировать статистическую базу данных, используемую в дальнейшем для решения целого ряда прикладных задач. Большинство достижений в статистике сначала появляются в виде пакетов R, и лишь затем добавляются в меню программных пакетов. Особенно плодотворной является интеграция SPSS, *STATISTICA* и R, позволяющая расширить возможности данных сред, использовать библиотеки и программы R внутри данных статистических пакетов.

Ключевые слова: язык R, база данных, статистический анализ, регрессионный анализ.

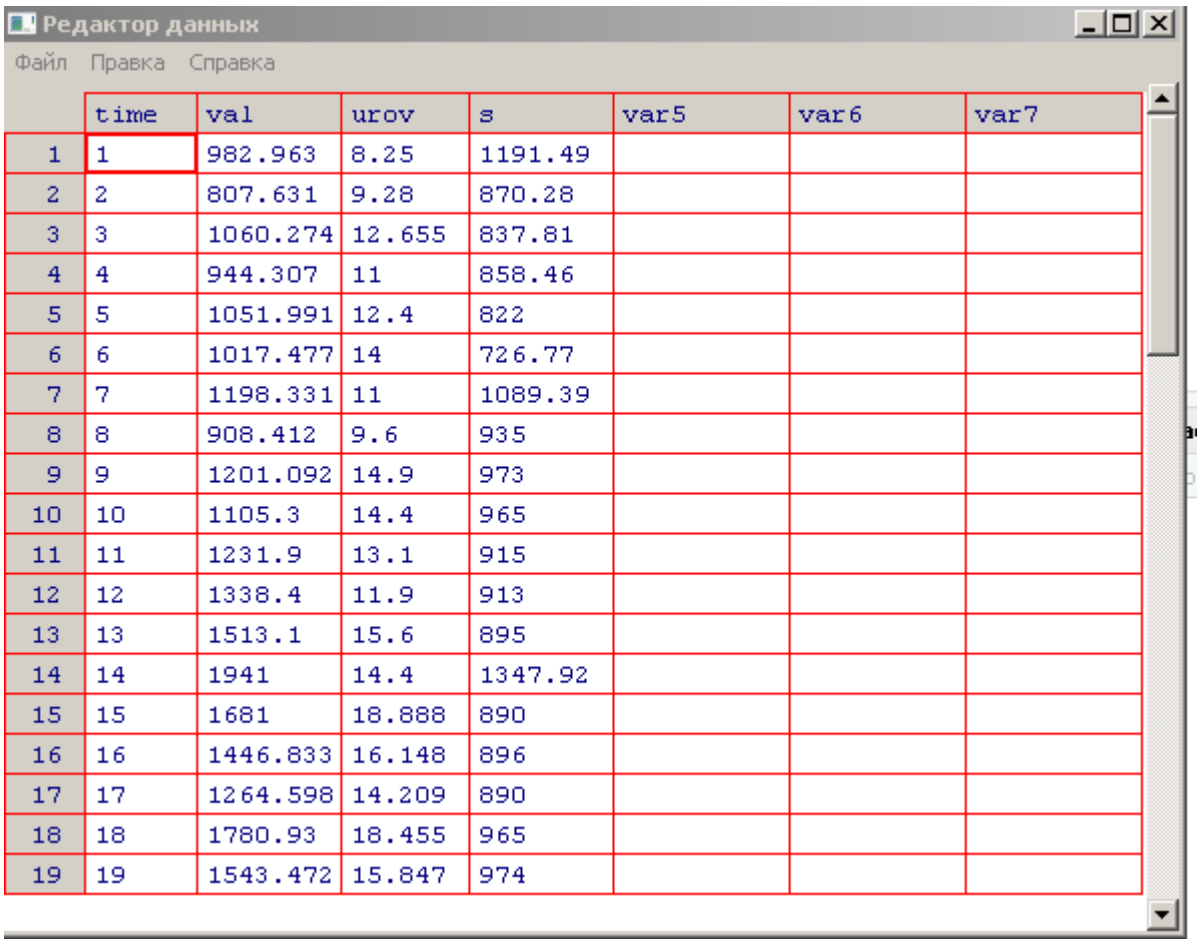
Первый этап любого анализа данных – создание набора данных, в котором содержится информация для изучения, в подходящем формате. Данные можно вводить вручную или импортировать из внешнего источника. Таким источником могут быть текстовые файлы, электронные таблицы, статистические программы и системы управления базами данных. Обычно исследователь сталкивается с данными, которые поступают из разных источников и в разных форматах. Задача состоит в том, чтобы импортировать данные в программу, проанализировать их и представить отчет о результатах. В R реализованы разные способы импорта данных с различной структурой, включая скаляры, векторы, массивы данных, таблицы данных и списки [1, 2, 3].

Наиболее широко используемым объектом является таблица данных (data.frame), в котором разные столбцы могут содержать разные типы данных. Каждый столбец должен содержать данные только одного типа, при этом в одной таблице данных могут быть столбцы с данными разного типа.

Нами был создан файл lipobl (типа data.frame) по сельскохозяйственным организациям Липецкой области, содержащий данные по зерновому производству: time – время в годах; urov – урожайность зерновых (ц/га); val – валовое производство зерна (тыс. тонн); s – площадь, отведенная под посев зерновых культур.

В начале создается пустая таблица данных (или матрица), с указанием названий и типов переменных. В нашем случае всем переменным был присвоен числовой тип (numeric). Затем с помощью функции edit() открываем текстовый редактор, куда вносим свои данные и сохраняем результат в виде объекта с данными [1, 4]:

```
lipobl <- data.frame(time=numeric(),urov=numeric(),val=numeric(),s=numeric())
> lipobl <- edit(lipobl)
```



	time	val	urov	s	var5	var6	var7
1	1	982.963	8.25	1191.49			
2	2	807.631	9.28	870.28			
3	3	1060.274	12.655	837.81			
4	4	944.307	11	858.46			
5	5	1051.991	12.4	822			
6	6	1017.477	14	726.77			
7	7	1198.331	11	1089.39			
8	8	908.412	9.6	935			
9	9	1201.092	14.9	973			
10	10	1105.3	14.4	965			
11	11	1231.9	13.1	915			
12	12	1338.4	11.9	913			
13	13	1513.1	15.6	895			
14	14	1941	14.4	1347.92			
15	15	1681	18.888	890			
16	16	1446.833	16.148	896			
17	17	1264.598	14.209	890			
18	18	1780.93	18.455	965			
19	19	1543.472	15.847	974			

Рисунок 1 – Создание таблицы данных data.frame

Присвоения типа time=numeric() создают пустую (без данных) переменную заданного типа. Функция edit() работает с копией объекта. Результат работы функции edit() под Windows показан на рис. 1.

Щелкая по названиям столбцов, можно изменить название и тип соответствующей переменной. Можно добавлять дополнительные переменные, щелкая на названия неиспользованных столбцов. После того как закрывается текстовый редактор, результаты сохраняются в виде выбранного объекта (в данном случае объект lipobl). Повторное введение функции lipobl <-edit(lipobl) позволяет редактировать введенные данные и добавлять новые.

Используя RCommander можно провести анализ таблицы данных data.frame (ниже приведена корреляционная матрица файла lipobl)

```
Rcmdr>  cor(lipobl[,c("s","time","urov","val")], use="complete")
             s       time       urov        val
s     1.0000000 -0.6043163 -0.4882067 -0.1085360
time -0.6043163  1.0000000  0.8542785  0.6763775
urov -0.4882067  0.8542785  1.0000000  0.9013877
val  -0.1085360  0.6763775  0.9013877  1.0000000
```

```
Rcmdr>  with(lipobl, lineplot(time, urov))
```

И график временного ряда изменчивости урожайности зерновых культур

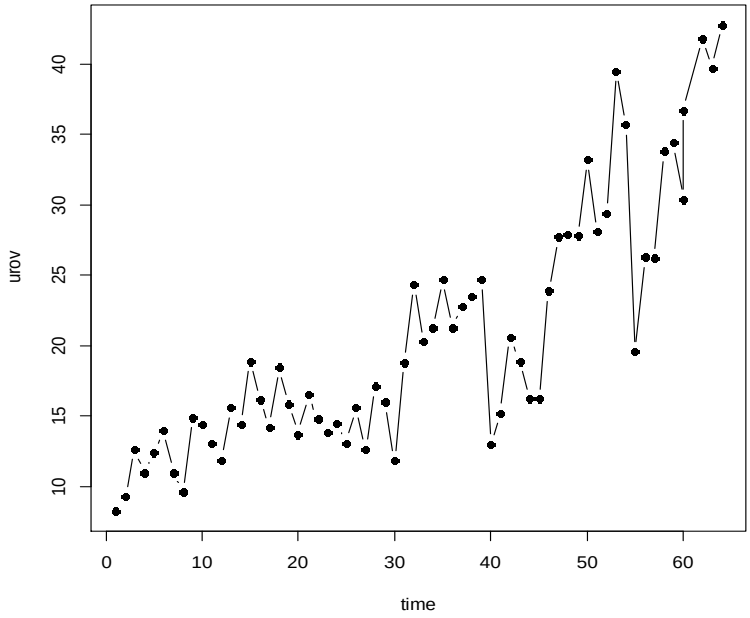


Рисунок 2 – График временного ряда урожайности зерновых культур

Язык R позволяет существенно упростить процедуру построения нелинейной множественной регрессии. В частности, нами было показано, что производственная функция, описывающая зависимость между объемом производимой продукции и величиной затраченных ресурсов, носит принципиально нелинейный характер [3, 5, 6].

В наших исследованиях чаще всего мы использовали кинетическую производственную функцию:

$$Y = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdot \ldots \cdot x_n^{\alpha_n} \cdot e^{a_1 x_1 + a_2 x_2 + \ldots + a_n x_n}$$

где Y – объем производимой продукции, $x_j \left(j = \overline{1,n} \right)$ – затраты ресурсов.

Для определения ее параметров прологарифмируем обе части данного выражения:

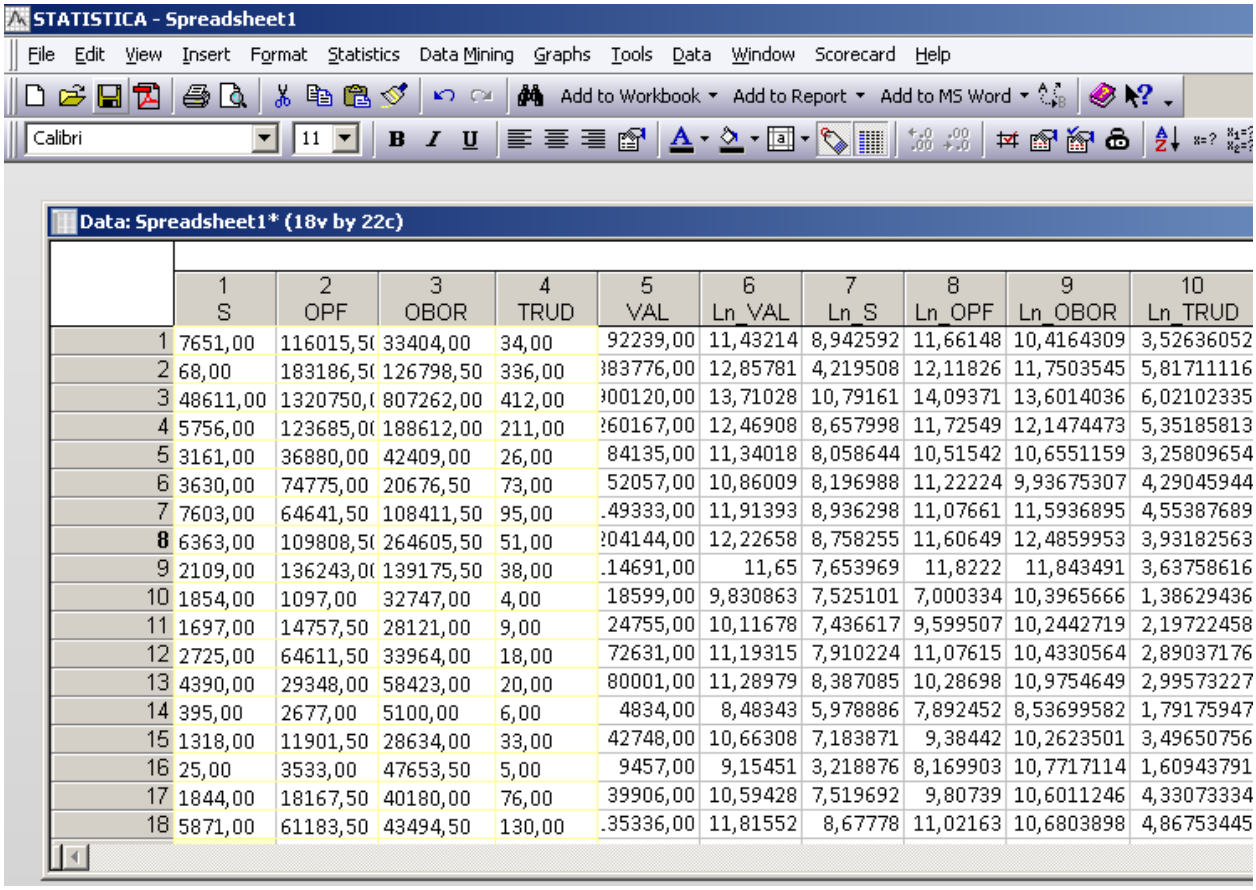
$$\ln Y = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n + \alpha_1 \ln x_1 + \alpha_2 \ln x_2 + \ldots + \alpha_n \ln x_n$$

Нами проводился расчет кинетической производственной функции с включением следующих переменных: Y – объём валовой продукции, x_1 – площадь сельскохозяйственных угодий, га; x_2 – основные производственные

фонды, тыс. руб., x_3 – размер оборотных средств, тыс. руб.; x_4 – количество работников, человек.

Для расчета данной производственной функции кроме вышеперечисленных переменных требуется создать еще дополнительные переменные: lnY, lnx_1, lnx_2, lnx_3, lnx_4.

В нижеприведенном рисунке приведен фрагмент исходной таблицы данных (Spreadsheet) в программе STATISTICA, в которой именами S, OPF, OBOR, TRUD, VAL обозначены соответственно площадь сельскохозяйственных угодий, га; основные производственные фонды, тыс. руб., размер оборотных средств, тыс. руб.; количество работников, человек. Далее выделены еще пять переменных (пять столбцов) для введения логарифмов указанных величин [7-9].



	1 S	2 OPF	3 OBOR	4 TRUD	5 VAL	6 Ln_VAL	7 Ln_S	8 Ln_OPF	9 Ln_OBOR	10 Ln_TRUD
1	7651,00	116015,5(	33404,00	34,00	92239,00	11,43214	8,942592	11,66148	10,4164309	3,52636052
2	68,00	183186,5(	126798,50	336,00	383776,00	12,85781	4,219508	12,11826	11,7503545	5,81711116
3	48611,00	1320750,(	807262,00	412,00	000120,00	13,71028	10,79161	14,09371	13,6014036	6,02102335
4	5756,00	123685,0(	188612,00	211,00	260167,00	12,46908	8,657998	11,72549	12,1474473	5,35185813
5	3161,00	36880,00	42409,00	26,00	84135,00	11,34018	8,058644	10,51542	10,6551159	3,25809654
6	3630,00	74775,00	20676,50	73,00	52057,00	10,86009	8,196988	11,22224	9,93675307	4,29045944
7	7603,00	64641,50	108411,50	95,00	.49333,00	11,91393	8,936298	11,07661	11,5936895	4,55387689
8	6363,00	109808,5(	264605,50	51,00	204144,00	12,22658	8,758255	11,60649	12,4859953	3,93182563
9	2109,00	136243,0(	139175,50	38,00	.14691,00	11,65	7,653969	11,8222	11,843491	3,63758616
10	1854,00	1097,00	32747,00	4,00	18599,00	9,830863	7,525101	7,000334	10,3965666	1,38629436
11	1697,00	14757,50	28121,00	9,00	24755,00	10,11678	7,436617	9,599507	10,2442719	2,19722458
12	2725,00	64611,50	33964,00	18,00	72631,00	11,19315	7,910224	11,07615	10,4330564	2,89037176
13	4390,00	29348,00	58423,00	20,00	80001,00	11,28979	8,387085	10,28698	10,9754649	2,99573227
14	395,00	2677,00	5100,00	6,00	4834,00	8,48343	5,978886	7,892452	8,53699582	1,79175947
15	1318,00	11901,50	28634,00	33,00	42748,00	10,66308	7,183871	9,38442	10,2623501	3,49650756
16	25,00	3533,00	47653,50	5,00	9457,00	9,15451	3,218876	8,169903	10,7717114	1,60943791
17	1844,00	18167,50	40180,00	76,00	39906,00	10,59428	7,519692	9,80739	10,6011246	4,33073334
18	5871,00	61183,50	43494,50	130,00	.35336,00	11,81552	8,67778	11,02163	10,6803898	4,86753445

Рисунок 3 – Фрагмент исходной таблицы данных (Spreadsheet)

При использовании языка R оценивание линейной регрессии методом наименьших квадратов осуществляется с использованием функции

lm(formula, data, subset, weights, ...)

lm = Linear Model – линейная модель. Напомним, что под линейной моделью понимается модель линейная по параметрам, так что после логарифмирования кинетическая производственная функция становится моделью линейной по параметрам.

Таблица 1

Основные аргументы [8, 10]

formula	спецификация регрессии (объект класса formula)
data	data.fraim фрейм, на котором оценивается модель
subset	(опционально) используется, если нужно оценить модель не по всему data.fraim, а только по его части
weights	(опционально) вектор весов для метода WLS

В результате выполнения функции создается объект типа lm, который представляет собой список, содержащий информацию о подогнанной модели. Для вывода протокола оценивания можно использовать функцию **summary().**

Если зависимая переменная или регрессоры в модели берутся с логарифмами (как в нашем случае), то в спецификацию модели нужно вставить функцию log():

$$\log(y) \sim x1 + x2 + x3 + x4 + \log(x1) + \log(x2) + \log(x3) + \log(x4)$$

и нет необходимости вводить столбцы с логарифмами.

Особо следует отметить тот факт, что большинство достижений в статистике сначала появляются в виде пакетов R, и лишь затем добавляются в меню программных пакетов. В значительной мере в силу этого интеграция языка R и наиболее значимых статистических пакетов (например, SPSS и *STATISTICA*) позволяет расширить возможности данных сред, и использовать программы R внутри данных статистических пакетов. Имеется возможность запускать библиотеки и скрипты R в оболочках SPSS и *STATISTICA*, получая результаты в виде отчетов, рабочих книг и графиков.

Список литературы:

1. Кабаков, Р.И. R в действии. Анализ и визуализация данных в программе R /Р.И. Кабаков. – М.: ДМК Пресс, 2014. – 588 с.

2. Смагин, Б.И. Логика формирования производственных функций/Б.И. Смагин, А.Б. Смагина//Развитие агропродовольственного комплекса: экономика, моделирование и информационное обеспечение: Сборник научных трудов. Воронеж, Воронежский ГАУ, 2016. – С. 97 – 105.

3. Уикем, Х. Язык R в задачах науки о данных: импорт, подготовка, обработка, визуализация и моделирование данных /Х. Уикем, Г. Гроулмунд. – СПб.: ООО «Диалектика», 2019. – 592с.

4. Смагин, Б.И. Определение производственного потенциала в аграрном производстве / Б.И. Смагин // Аграрная наука. - 2003. - № 1. - С. 4.

5. Смагин, Б.И. Предпрогноз временного ряда (на примере зернового производства в регионе) / Б.И. Смагин // В сб.: Субрегиональное сотрудничество в современных условиях развития национальной экономики: сборник научных трудов. – Воронеж: ООО «Издательство Ритм», 2020. – С. 49-52.

6. Смагин, Б.И. Анализ основных предположений, используемых при построении производственных функций / Б.И. Смагин // В сб.: Современный менеджмент: теория, методология и практика: материалы региональной научно-практической конференции, посвященной памяти д.э.н., профессора Т.К. Абдуллаевой. – Махачкала: Информационно-Полиграфический Центр ДГТУ, 2019. – С. 141-144.

7. Смагин, Б.И. Экономико-математический подход к оценке земель сельскохозяйственного назначения / Б.И. Смагин, А.Б. Смагина // В сб.: Современный менеджмент: теория, методология и практика: материалы региональной научно-практической конференции, посвященной памяти д.э.н., профессора Т.К. Абдуллаевой. – Махачкала: Информационно-Полиграфический Центр ДГТУ, 2019. – С. 308-311.

8. Смагин, Б.И. Алгоритм построения производственных функций в аграрной сфере производства / Б.И. Смагин // В сб.: Агротехнологии XXI века: материалы Всероссийской научно-практической конференции с

международным участием, посвященной 100-летию высшего аграрного образования на Урале. – Пермь: ИПЦ Прокростъ, 2019. – С. 97-100.

9. Смагин, Б.И. Математическое описание различных способов начисления амортизационных отчислений на основные производственные фонды / Б.И. Смагин // Наука и Образование. – 2019. – Т. 2. – № 4. – С. 216. 0

10. Plant protection and foliar fertilizing technology of apple (Malus domestica Borkh) / A.I. Kuzin, N.Ya. Kashirskaya, A.M. Kochkina, B.I. Smagin // International Journal of Engineering and Advanced Technology. - 2019. - Т. 8. - № 6. - С. 3613-3620.

UDC 681.3.07

SOME ASPECTS OF STATISTICAL ANALYSIS USING THE ALGORITHMIC LANGUAGE R

Smagin Boris Ignatievich
Doctor of Economic Sciences, Professor
bismagin@mail.ru
Michurinsk State Agrarian University
Michurinsk, Russia

Annotation. R is a programming language for statistical data processing and working with graphics, as well as a free software environment with open source code that allows you to create a statistical database that will be used in the future to solve a number of applied problems. Most achievements in statistics first appear as R packages before being added to the software package menu. The integration of SPSS, STATISTICA, and R is particularly fruitful, allowing you to expand the capabilities of these environments and use R libraries and programs inside statistical data sets.

Key words: R language, database, statistical analysis, regression analysis.