АЛГОРИТМ СТАТИСТИЧЕСКОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ МОДЕЛИ МНОГОМЕРНОЙ ДИНАМИЧЕСКОЙ РЕГРЕССИИ

В. И. Малюгин

Белорусский государственный университет,
Минск, Республика Беларусь
E-mail: Malugin@bsu.by

Для модели динамической многомерной линейной регрессии с ошибками наблюдений, описываемыми стационарной моделью векторной авторегрессии, предлагается итерационный двухшаговый алгоритм вычисления оценок параметров в соответствии с обобщенным методом наименьших квадратов (МНК). Приводятся результаты экспериментального исследования алгоритма на модельных данных.

Ключевые слова: динамической модель многомерной линейной регрессии. обобщенные МНК-оценки параметров, итерационный алгоритм.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ И ПОСТАНОВКА ЗАДАЧИ

Будем предполагать, что для произвольного момента времени t ($t=1,...T$) наблюдаемый объект описывается случайным вектором наблюдений $\mathbf{y}_t \in \Re^{N+M}$, который допускает разбиение на подвекторы:

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{pmatrix} \in \Re^{N+M}, \text{ где } \mathbf{x}_t \in \Re^N,\ \mathbf{z}_t \in \mathbf{Z} \subset \Re^M, \quad (1)$$

где $\mathbf{x}_t = (x_{t1},...,x_{tN})'$ – вектор эндогенных переменных (признаков), характеризующих состояние анализируемого объекта; $\mathbf{z}_t = (z_{t1},...,z_{tM})'$ – вектор экзогенных переменных (факторов) описывающих внешние воздействия на состояние объектов.

Между подвекторами составного вектора наблюдений (1) существует статистическая зависимость, которая описывается моделью многомерной линейной регрессии вида:

$$\mathbf{x}_t = B\mathbf{z}_t + \mathbf{v}_t,, \quad (2)$$

где $B=(b_{kl})$ – $(N\times M)$-матрица коэффициентов регрессии; $\mathbf{v}_t = (v_{t1},...,v_{tN})' \in \Re^N$ – вектор случайных ошибок наблюдения, который описывается моделью векторной авторегрессии первого порядка (*vector autoregressive* – *VAR*(1)) вида:

$$\mathbf{v}_t = A\mathbf{v}_{t-1} + \mathbf{u}_t, \quad (3)$$

для которой $A=(a_{kl})$ – $(N\times N)$-матрица коэффициентов авторегрессии; $\mathbf{v}_0 \in \Re^N$ – начальное значение случайного процесса $\mathbf{v}_t$.

Относительно модели (2)–(4) делаются следующие предположения.

П1. Случайный процесс $\mathbf{u}_t \in \Re^N$ является гауссовским белым шумом с нулевым средним значением и невырожденной ковариационной $(N \times N)$ – матрицей $\Sigma = (\sigma_{kl}) : E\{\mathbf{u}_t\} = 0,\ E\{\mathbf{u}_t\mathbf{u}'_\tau\} = \delta_{t\tau}\Sigma,$ где $\delta_{t\tau}$ – символ Кронекера; будем обозначать: $\{\mathbf{u}_t\}\sim$ н.о.р.с.в. $\mathbf{N}_N(\mathbf{0}, \Sigma)$.

П2. $\mathbf{z}_t \in \mathbf{Z} \subset \Re^M$, где $\mathbf{Z} = \{\mathbf{z} \in \Re^M : \mathbf{z}'\mathbf{z} \le c,\ 0 < c < \infty\}$ – ограниченная область.

П3. Модель векторной авторегрессии (3) удовлетворяет условию стационарности, т.е. корни алгебраического уравнения $\left|I_N\lambda - A_t\right| = 0$ лежат внутри единичного круга: $\left|\lambda_j\right| < 1,\ j = 1,\dots,N$, где $I_N - (N \times N)$ - единичная матрица.

Истинные значения параметров $A,\ B,\ \Sigma$, модели (2), (3) не известны. Имеется выборка значений признаков $X = \{\mathbf{x}_1,\dots,\mathbf{x}_T\}$ объема T, соответствующая последовательности значений факторов $Z = \{\mathbf{z}_1,\dots,\mathbf{z}_T\}$, где значение вектора признаков $\mathbf{x}_t \in \Re^N$ определяется на основании (3) для заданного значения вектора факторов $\mathbf{z}_t \in \mathbf{Z}$ $(t = 1,\dots,T)$.

Задача заключается в построении и исследовании точности статистических оценок параметров $A,\ B,\ \Sigma$, модели (2), (3) по реализациям временных рядов $X = \{\mathbf{x}_1,\dots,\mathbf{x}_T\}$, $Z = \{\mathbf{z}_1,\dots,\mathbf{z}_T\}$.

АЛГОРИТМ СТАТИСТИЧЕСКОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ

При N=1 модель (2), (3) представляет собой модель множественной линейной регрессии с автокоррелированными ошибками, для оценивания параметров которой традиционно используется обобщенный метод наименьших квадратов (МНК). Для практической реализации данного метода могут использоваться итерационные алгоритмы Кохрейна – Окартта, Хилдрета – Лу, Дарбина [1]. Ни один из этих подходов не может непосредственно применяться в рассматриваемом многомерном случае, когда N>1.

В соответствии с общей методологией построения обобщенных МНК-оценок осуществим преобразование переменных в модели (2), (3), которое позволит перейти от исходной модели к традиционной модели с некоррелированными случайными ошибками наблюдения. Обычные МНК-оценки параметров для преобразованной модели являются обобщенными МНК-оценками для исходной модели [1].

Лемма. Пусть для модели VAR(1) вида (3) выполняется условие стационарности П.3, $\mathbf{v}_0 \in \Re^N$ – заданное начальное значение, а случайный процесс $\{\mathbf{u}_t\}$ является бесконечным в обе стороны ($t = \dots -2,-1,0,1,2\dots$) и удовлетворяет предположению П.1, тогда модель (2), (3) допускает эквивалентное представление в виде модели VAR(1) с экзогенными переменными и некоррелированными ошибками:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{z}_t - AB\mathbf{z}_{t-1} + \mathbf{u}_t = A\mathbf{x}_{t-1} + C\mathbf{Z}_t + \mathbf{u}_t, \qquad (4)$$

где: $C = (C_0|C_1)$ – блочная $(N \times 2M)$-матрица $(C_0 = B, C_1 = -AB)$; $\mathbf{Z}'_t = (\mathbf{z}'_t;\mathbf{z}'_{t-1}) \in \mathbf{Z}^{2M}$ – составной вектор факторов.

Следствие. Если в модели (2), (3) осуществить преобразование переменных вида

$$\mathbf{y}_t = \mathbf{x}_t - A\mathbf{v}_{t-1}, \qquad (5)$$

то в условиях леммы модель (2), (3) допускает представление в виде модели многомерной линейной регрессии с некоррелированными ошибками:

$$\mathbf{y}_t = B\mathbf{z}_t + \mathbf{u}_t. \qquad (6)$$

В предположениях П.1–П.3 по аналогии с [2] могут быть получены аналитические представления для оценок параметров A, C и Σ модели (4) по реализации временного ряда значений зависимых переменных $\{\mathbf{x}_t\}$, соответствующей заданной реализации временного ряда значений факторов $\{\mathbf{z}_t\}$ $(t = 1,...T)$. Однако, получаемые в данном случае аналитические формулы не удобны в вычислительном отношении и «избыточно сложны» для вычисления параметров A, B и Σ модели (2), (3): требуется оценивать $2MN+N^2+N^2$ параметров (элементов матриц A, C и Σ), чтобы найти $MN+N^2+N^2$ искомых параметров (элементов матриц A, B, и Σ). При небольшом объеме данных это отражается на точности оценок. По этой причине приведем описание алгоритма вычисления обобщенных МНК-оценок параметров модели (2), (3), который может быть реализован, как автономно, так и с использованием статистических (эконометрических) пакетов прикладных программ, имеющих средства оценивания векторных авторегрессионных моделей с экзогенными переменными при выполнении традиционных предположений о некоррелированности случайных ошибок наблюдения (см. обзор в [3]). Последняя возможность является принципиально важной, поскольку, как известно, оцениванием параметров моделей статистических зависимостей не исчерпываются проблемы, связанные с их построением. Важным этапом в построении таких моделей является этап анализа статистической адекватности оцененной модели на основе статистических критериев и тестовых статистик. При реализации предлагаемого алгоритма с помощью статистических (эконометрических) пакетов прикладных программ исследователь может использовать все имеющиеся возможности пакета, как для вычисления оценок параметров, так и для анализа адекватности построенной модели.

Предлагаемый алгоритм следует схеме алгоритма Кохрейна – Окартта в случае N=1, однако использует для исходных переменных в многомерной модели (2), (3) другой вариант преобразования (преобразования (5)), приводящих данную модель к виду (6).

Описание алгоритма. Для обозначения номера итерации будем использовать верхний индекс k в скобках при соответствующих параметрах и переменных. Опишем действия, которые выполняются на начальной (k=0) и последующих (k=1, 2,...) итерациях алгоритма.

Начальная итерация (k=0)

Шаг 0.1. Полагаем: $A^{(0)} = O_N$ (т.е. считаем случайные ошибки $\{\mathbf{v}_t\}(t = 1,...,T)$ некоррелированными); $B^{(0)}$ – МНК-оценка матрицы B в модели

$$\mathbf{x}_t = B\mathbf{z}_t + \mathbf{v}_t,\ t = 1,...,T. \qquad (7)$$

Шаг 0.2. Вычисляем остатки для модели (7) по формуле

$$\mathbf{v}_t^{(0)} = \mathbf{x}_t - B^{(0)}\mathbf{z}_t,\ t = 1,...,T.$$

Итерация k (k=1, 2, ...)

Шаг k.1. Полагаем $A^{(k)}$ – МНК-оценка матрицы A в модели

$$\mathbf{v}_t^{(k-1)} = A\mathbf{v}_{t-1}^{(k-1)} + \mathbf{u}_t.$$

Шаг k.2. Проверяем условие остановки алгоритма. Если элементы матриц $A^{(l)} = (a_{ij}^{(l)})$ $(l = k-1, k)$ удовлетворяют условию:

$$\left| \left\| A^{(k)} \right\| - \left\| A^{(k-1)} \right\| \right| \le \varepsilon,\ 0 < \varepsilon \ll 1,$$

то полагаем: $\hat{A} = A^{(k-1)}$, $\hat{B} = B^{(k-1)}$ и работа алгоритма прекращается. Здесь $\|A\| = \sqrt{\sum_{kl}^{N} a_{kl}^2}$ – норма матрицы $A = (a_{kc})$. В противном случае переходим к шагу k.2.

Шаг k.2. Осуществляем преобразование переменных по формуле

$$\mathbf{y}_t^{(k)} = \mathbf{x}_t - A^{(k)}\mathbf{v}_{t-1}^{(k-1)},\ t = 1,...,T. \tag{8}$$

Полагаем $B^{(k)}$ – МНК-оценка матрицы B в модели

$$\mathbf{y}_t^{(k)} = B\mathbf{z}_t + \mathbf{v}_t,\ t = 1,...,T. \tag{9}$$

Шаг k.3. Вычисляем остатки для модели (9) по формуле

$$\mathbf{v}_t^{(k)} = \mathbf{x}_t - B^{(k)}\mathbf{z}_t,\ t = 1,...,T,$$

полагаем $k := k+1$ и переходим к шагу k.1 на новой итерации.

С учетом специфики задачи статистического оценивания параметров, в качестве характеристики сложности алгоритма целесообразно использовать общее число оцениваемых параметров модели по выборке фиксированного объема. По этой характеристике обычный МНК (число оцениваемых параметров равно $NM+N^2$ – общее число элементов матриц B и Σ соответственно) превосходит предлагаемый алгоритм ($NM+N^2+N^2$ – общее число элементов матриц B, A и Σ) и, тем более, алгоритм вычисления параметров по аналитическим формулам ($2MN+N^2+N^2$ – общее число элементов матриц C, A, и Σ). Однако, обычный МНК не учитывает автокорреляцию случайных ошибок наблюдений (матрица A не оценивается) и поэтому следует ожидать выигрыш в точности оценивания параметров при использовании специально разработанного итерационного алгоритма.

Сложность рассматриваемой задачи как задачи статистического оценивания параметров для обычного МНК можно характеризовать показателем $\delta = \|A\| / N \ge 0$. При $\delta = 0$ выполняется традиционное для МНК предположение о некоррелированности ошибок наблюдения. Чем больше значение δ, тем сильнее степень нарушения этого предположения.

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМА

В качестве иллюстрации работоспособности алгоритма проведем результаты двух серий экспериментов для различных тестовых моделей, преследующих следующие цели:

1) сравнительный анализ точности оценок параметров модели (2), (3) при различных значениях δ на основе предлагаемого алгоритма, обычного МНК и альтернативных алгоритмов вычисления обобщенных МНК-оценок (алгоритмов Кохрейна – Окартта и Дарбина [1]) для модели множественной линейной регрессии (N=1) с ошибками наблюдений, описываемыми моделью авторегрессии первого порядка AR(1) т.е. в условиях, когда возможно применение альтернативных алгоритмов;

2) сравнительный анализ точности оценок параметров модели многомерной модели (2), (3), получаемых с помощью обычного МНК и предлагаемого алгоритма при различных значениях δ, соответствующих различной степени зависимости регрессионных наблюдений.

Тестовая модель 1 определяется следующими соотношениями: $N = M = 1$, $T = 12$,

$$x_{t1} = b_1 z_t + v_{t1},\ v_{t1} = a_{11} v_{t-1,1} + u_{t1},\ u_{t1} \sim \text{н.о.р.с.в. } \mathbf{N}(0, \sigma_{11}^2),$$

значения z_t равномерно распределены на интервале $[-1, 1]$, $a_{11} \in \{0.5, 0.9, 0.95\}$, $\delta = a_{11}$, $b_1 = 2$, $\sigma_{11} = 0.625$. Оценки параметров для рассматриваемых алгоритмов (усредненные по 5 прогонам) представлены в табл. 1.

В случае алгоритма Дарбина оцениваются параметры модели вида

$$x_t = a_{11} x_{t-1} + b_1 \mathbf{z}_t - a_{11} b_1 z_{t-1} + u_t = a_{11} x_{t-1} + b_1 z_t + d z_{t-1} + u_t,$$

(частного случая модели (4) при N=1) поэтому в таблицах указываются две оценки параметра b_1: оценка коэффициента регрессии при переменной $\mathbf{z}_t$ и оценка, вычисляемая как $\hat{d} / \hat{a}_{11}$.

В столбце МНК для параметра a_{11} приводятся МНК-оценки по смоделированной реализации случайных ошибок $\{v_{t1}\}$.

Таблица 1

Модель множественной линейной регрессии с ошибками вида AR(1) (T=12)

Истинное значение a_{11}	Параметры	МНК	Алгоритм Кохрейна – Окаргта	Алгоритм Дарбина	Предлагаемый алгоритм
0.5	a_{11}	0.3041	0.2278	0.2555	0.2278
	b_1	2.0746	2.0619	2.1238/1.3102	2.1050
0.9	a_{11}	0.8705	0.8502	0.8379	0.8502
	b_1	2.2027	1.9895	2.1504/1.7536	2.1129
0.95	a_{11}	0.9497	0.9307	0.9258	0.9307
	b_1	2.3407	1.9781	2.1436/1.7886	2.0958

Тестовая модель 2 определяется следующими соотношениями: $N = 2$, $M = 1$, $T = 300$,

$$x_{t1} = b_1 z_t + v_{t1}, x_{t2} = b_2 z_t + v_{t2},\ v_{t1} = a_{11} v_{t-1,1} + a_{12} v_{t-1,2} + u_{t1},\ v_{t2} = a_{21} v_{t-1,1} + a_{22} v_{t-1,2} + u_{t2},$$

$u_{tj} \sim$ н.о.р.с.в. $\mathbf{N}(0, \sigma_j^2)(j = 1, 2)$, значения z_t равномерно распределены на интервале $[-1, 1]$.

Рассматривалось два тестовых примера, отличающихся степенью корреляции случайных ошибок наблюдения в модели (2), (3):

1) $b_1 = 2, b_1 = 5, a_{11} = 0.4, a_{12} = 0, a_{21} = -0.2, a_{22} = 0.5, \sigma_1 = 0.8, \sigma_2 = 0.4$, δ=0.5;

2) $b_1 = 2, b_1 = 5, a_{11} = 0.9, a_{12} = 0, a_{21} = -0.7, a_{22} = 0.8, \sigma_1 = 0.8, \sigma_2 = 0.4$, δ=0.7.

Оценки параметров моделей для рассматриваемых случаев на основе обычного МНК, не учитывающего автокорреляцию случайных ошибок, и предлагаемого алгоритма (усредненные по 5 прогонам) представлены в табл. 2.

Таблица 2

Модель многомерной линейной регрессии с ошибками вида VAR(1)

Тестовый пример	Параметры	МНК	Предлагаемый алгоритм (ПА)
1	β_1, β_2	1.9948, 5.0009	1.9981, 4.9258
2	β_1, β_2	1.8717, 5.6312	1.9906, 4.9486

В табл. 3 представлены значения основных тестовых статистик, используемых для оценки качества построенных многомерных статистических моделей, включая [1, 3]: информационные статистики Акаике (AIC) и Шварца (SC), нормированную матричную сумму квадратов остатков $\left|\hat{\Sigma}\right|$ и логарифмическую функцию правдоподобия для вычисленных оценок $l_N\left(\hat{A},\hat{C},\hat{\Sigma}\right)$. Меньшие значения статистик AIC, SC, $\left|\hat{\Sigma}\right|$ и большие значения статистики $l_N\left(\hat{A},\hat{C},\hat{\Sigma}\right)$ соответствуют более качественной модели. Заметим также, что для всех моделей оценки коэффициентов регрессии являются статистически значимыми на уровне 0.05, а для моделей, построенных с помощью предлагаемого алгоритма, остатки являются гауссовским белым шумом.

Таблица3

Характеристики качества статистических моделей

Алгоритм	Тестовые статистики			
	AIC	SC	$\left\|\hat{\Sigma}\right\|$	$l_N\left(\hat{A},\hat{C},\hat{\Sigma}\right)$
		Тестовый пример 1		
МНК	4.0736,	4.0983,	0.2001,	-609.045
ПА	3.9312	3.9556	0.1736	-585.714
		Тестовый пример 2		
МНК	8.1854	8.2101	12.2190	-1225.810
ПА	3.4506	3.4754	0.1073	-513.8673

На основании проведенных экспериментов можно сделать следующие выводы.

1. В случае N=1 (табл. 1) предлагаемый алгоритм по точности оценивания регрессионных коэффициентов занимает промежуточное положение между алгоритмами Кохрейна – Окартта и Дарбина; можно также показать, что преимущество итерационных алгоритмов над обычным МНК проявляется в условиях коротких временных рядов и тем сильнее, чем сильнее степень зависимости регрессионных наблюдений.

2. В случае N>1 при высокой степени зависимости регрессионных наблюдений предлагаемый алгоритм имеет существенный выигрыш по сравнению с обычным МНК как по точности оценивания регрессионных коэффициентов (табл. 2), так и по всем основным характеристикам качества статистических моделей (табл.3).

ЛИТЕРАТУРА

1. *Магнус, Я. Р.* Эконометрика. Начальный курс / Я. Р., Магнус, П. К. Катышев, А. А. Пересецкий. М.: Дело, 2004. 576 с.

2. *Андерсон, Т.* Статистический анализ временных рядов / Андерсон Т. М.: Мир, 1976. 755 с.

3. *Харин, Ю.С.* Эконометрическое моделирование / Ю. С. Харин, В. И. Малюгин, А. Ю. Харин. Мн.: БГУ, 2003. 313 с.