Кадиров Р.Х.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ ПРИ ИССЛЕДОВАНИИ ЖИВЫХ СИСТЕМ (ПРОБЛЕМА СЕЗОННЫХ КОЛЕБАНИЙ)

Статья посвящена проблеме сезонных колебаний процессов, протекающих в живых системах. Эта проблема исследована на примере метода наименьших квадратов, и предложен подход, основанный на включении параметров внешней среды для учета качественной компоненты временного фактора. Рассматриваются особенности применения метода наименьших квадратов при исследовании процессов, протекающих в живых системах. Проведен анализ причин возникновения неэффективных оценок параметров модели в условиях нарушения предположения об отсутствии одновременности между случайными возмущениями и матрицей наблюдений. Предлагаются подходы, позволяющие повысить эффективность оценок в условиях возникновения проблемы сезонных колебаний.

На современном этапе медико-биологические исследования носят комплексный, системный характер, охватывающий различные стороны исследуемого процесса. При этом отмечается значительный рост количества информации, что существенно усложняет ее анализ.

Так как организм описывается большим числом параметров и множеством внутренних взаимосвязей, то для его изучения часто используют метод наименьших квадратов (МНК). Это обусловлено тем обстоятельством, что на сегодняшний день МНК является всесторонне изученным и имеющим несколько теоретических обоснований [7, 11].

Главными преимуществами оценок МНК является то, что они обладают минимально возможной дисперсией в классе всех линейных несмещенных оценок и являются соответственно наилучшими линейными несмещенными оценками неизвестных параметров [3, 8].

По мере усложнения исследуемого процесса и вовлечения в рассмотрение новых компонентов оценки МНК можно использовать в качестве первого приближения [6, 12]. Все эти преимущества и являются причиной широкого использования МНК в медицине при исследовании различных патологических процессов [9].

Постановка задачи метода наименьших квадратов заключается в следующем. Пусть существует линейное соотношение между переменной Y, $k-1$ объясняющими переменными $X_2, X_3, ..., X_k$ и возмущением u. Если мы имеем выборку из n наблюдений над переменными Y и X_j, $j = 2,....,k;$ то можно записать:

$$Y_i = \sum_{i=1}^{k} \beta_i X_i + \beta_0 + u_i, \quad (i = 1,....,n) \qquad (1)$$

Коэффициенты β_k и параметры распределения u неизвестны. Задача состоит в получении эффективных оценок неизвестных коэффициентов β_k в соответствии с критерием минимизации: $E[Y_i - S]^2 \Rightarrow min$.

Здесь E – оператор математического ожидания, S – объясняемая переменная.

Для того чтобы оценить коэффициенты β_k, в методе наименьших квадратов принимаются предположения относительно того, как генерируются наблюдения. В связи с этим выдвигаются следующие основные гипотезы:

1) Предположение о нулевом среднем значении $E(u) = 0$.

При исследовании какого-либо процесса мы не можем охватить все параметры, имеющие влияние на исследуемый процесс, и включаем в рассмотрение, как правило, наиболее существенные. Эта гипотеза требует, чтобы суммарное влияние исключенных переменных на исследуемый процесс взаимно компенсировалось и сводилось к нулю.

2) Предположение о постоянстве условной дисперсии и об отсутствии автокорреляции $E(uu') = \Omega^2 I_n$.

Это предположение фактически требует, чтобы эксперимент был проведен при стабильных параметрах окружающей среды и чтобы влияние исключенных из рассмотрения переменных не преумножалось при их сочетании.

3) Предположение об отсутствии одновременности $E(X'u) = 0$ – фиксировано в повторяющихся выборках либо распределено независимо от u.

Эта гипотеза требует, чтобы рассматриваемые нами показатели не имели функциональных зависимостей с исключенными. Выполнение этого требования позволяет все случайные возмущения полученной модели отнести на счет влияния исключенных факторов, и тогда все свойства оценок β_k будут обусловлены качеством исходного вариационного ряда.

4) Предположение о ранге $\rho(X) = k < n$.

Это условие требует, чтобы между исследуемыми параметрами не было функциональных связей, то есть исключается линейная зависимость между ними.

На практике, при исследовании живых систем и процессов, протекающих в них, не всегда представляется возможным удовлетворить все эти требования. Это связано с тем, что живой организм – динамическая, развивающаяся система, характеризующаяся множеством внутренних и внешних связей различных систем организма и параметров внешней среды [5]. Так как значения исследуемых параметров меняются с течением времени, то есть динамичны по своей природе, то это приводит к нарушению требований 1, 2, 3 МНК. В результате возникают проблемы автокорреляции возмущений и корректировки сезонных колебаний [2, 4].

В случаях, когда между возмущениями имеется автокорреляция, но она не полностью коррелирована, использование обычного МНК позволит получить линейные и несмещенные оценки, но они не будут обладать наименьшей дисперсией. В таких ситуациях обычно применяют обобщенный МНК Эйткена [15].

Рассмотрим более подробно проблему корректировки сезонных колебаний. Именно от решения данной проблемы напрямую зависит решение ряда экологических задач [13]. Это обусловлено тем, что параметры окружающей среды оказывают существенное влияние на функционирование живых систем [1, 14]. Оптимальными, с точки зрения биологических объектов, считаются такие значения параметров внешней среды, при которых на адаптационные механизмы живых систем оказывается нагрузка, не превышающая их возможностей. Именно поэтому критические значения параметров внешней среды определяют в привязке к адаптационным возможностям организма.

На этом этапе возникает ряд проблем, требующих своего решения, в частности:

1) учет индивидуальных особенностей организма, определяющих адаптационные возможности каждого организма;

2) определение роли и степени влияния каждого параметра как внутренней, так и внешней среды на исследуемый процесс;

3) учет адаптационных возможностей живой системы в динамике.

Для учета индивидуальных особенностей организма нет другой альтернативы, кроме как создание более однородных групп, объединенных по схожести качественных признаков. При этом исходный массив данных разбивается на подмассивы по следующим критериям: пол, возраст, тип тела, тип нервной системы, группа крови, социальное положение, место жительства и так далее.

Для определения роли и места каждого из рассматриваемых параметров необходимо построение системы регрессионных уравнений. Для того чтобы весовые коэффициенты уравнений отражали истинную степень влияния параметра на исследуемый процесс, следует предварительно произвести нормировку данных. Аппарат математической статистики имеет в арсенале процедуры, позволяющие произвести нормировку данных на основе математического ожидания, среднеквадратического отклонения, минимального и максимального значений показателя.

При исследовании медицинских процессов все большее развитие приобретает нормировка, связанная со значением показателя, соответствующего норме [10]: $\overline{X}$: $X_i^* = X_i / \overline{X}$.

В результате исходная задача (1) преобразуется к виду:

$$Y_i = \sum_{i=1}^{k} \beta_i (X_i / \overline{\overline{X}}) + \beta_0 + u_i, \quad (i = 1,\ldots,n) \qquad (2)$$

Третья проблема – учет адаптационных возможностей живой системы в динамике фактически и составляет задачу корректировки сезонных колебаний. Она возникает вследствие того, что переменные исходного массива данных *X* порой сами являются функциями от некоторых других параметров. Эта зависимость может быть как линейной, так и нелинейной. Такие переменные называются фиктивными. Можно существенно расширить сферу применения линейных моделей, допустив включение в матрицу *X* фиктивных переменных. Как правило, эти переменные отражают изменения в таких факторах, каким является эффект сдвига во времени или в пространстве, воздействие качественных переменных или большого числа количественных переменных. При этом модель должна оставаться линейной относительно своих коэффициентов и удовлетворять всем свойствам, необходимым для применения обыкновенного метода наименьших квадратов.

Заметим здесь, что введение фиктивных переменных и применение обычной процедуры вычисления регрессии, которая автоматически определяет величины свободного члена, порой приводит к нарушению процесса оценивания.

Это обстоятельство возникает в случае, если введенная фиктивная переменная является линейной (или почти линейной) комбинацией остальных.

Фиктивные переменные играют важную роль в решении задачи корректировки сезонных колебаний. Эта задача сводится в следующим двум постановкам:

во-первых – ее можно рассматривать как традиционную проблему устранения сезонных возмущений в квартальных или месячных рядах;

во-вторых – это задача оценивания соотношений между переменными, среди которых имеются как скорректированные, так и подверженные влиянию сезонных воздействий. Решение этой постановки задачи, как правило, сводится к решению первой.

При оценивании соотношений между переменными, опираясь на сезонную статистику, в первую очередь необходимо ответить, не позволяет ли сама теория каким-либо образом получить данные с исключенной сезонностью или в нескорректированной форме. Это связано с тем, что во многих случаях нельзя точно определить, какие данные скорректированы, а какие – нет. А следовательно, неясно, какие переменные следует включать в регрессию. Однако на практике проблема спецификации оказывается менее существенной благодаря результатам [16]. Здесь рассмотрены две регрессии наименьших квадратов:

$$y = Xc_1 + Db_1 + e_1 \qquad (3)$$

$$y^{\alpha} = X^{\alpha}c_2 + e_2 \qquad (4)$$

Здесь матрицу D имеет порядок $4n*4$, а $4n$ – это количество сезонных наблюдений над переменной Y,

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (5)$$

$$y^{\alpha} = Ay \quad \text{и} \quad X^{\alpha} = AX \qquad (6)$$

и

$$A = I - D(D'D)^{-1}D' \qquad (7)$$

и доказано, что $c_1 = c_2$, то есть регрессии (3) и (4) дают одни и те же оценки коэффициентов при переменных X. При этом в (6) и (7) не накладывается никаких условий относительно элементов матрицы D (5), а лишь утверждается, что y и X скорректированы посредством регрессии на общую матрицу данных D, так что y^{α} и X^{α} являются остатками от этих регрессий. Также утверждается, что регрессии

$$y^{\alpha} = X^{\alpha}c_3 + e_3 \qquad (8)$$

$$y = X^{\alpha}c_4 + Db_4 + e_4 \qquad (9)$$

также определяют идентичный вектор коэффициентов при переменных X, то есть

$$c_1 = c_2 = c_3 = c_4 \qquad (10)$$

Таким образом, регрессии y на X и D (3), y на X^{α} (8) или y на X^{α} и D (9) во всех случаях дадут те же оценки коэффициентов при X, что и регрессия y^{α} на X^{α}.

Эти результаты приводят к новым вопросам. Действительно, если D – матрица только фиктивных переменных, то y^{α} и X^{α}, определенные в (6), строго говоря, не является скорректированными на сезонность рядами. С другой стороны, если должным образом скорректированные ряды получены с помощью матрицы преобразования A из (6), то эта матрица, хотя она и идемпотентная, не обладает свойством симметрии, требующимся для доказательства равенства c_1 и c_2. Поэтому интересно выяснить, какие различия между c_1 и c_2 появятся, если к должным образом скорректированным рядам применить регрессию (4).

Эти вопросы наводят на мысль, что подобные результаты имеют место в достаточном числе случаев [17]. Тем не менее мало кто предпочитает на практике воспользоваться этой теоремой. Вместо этого обычно аккуратно выбирают подходящую спецификацию соотношения, указывая при этом, какие переменные на какие должны реагировать непосредственно. Но произвол в выборе вида соотношения вносит в исследовательский процесс элемент субъективизма, который, вообще говоря, нежелателен, так как порой приводит в неверным выводам.

В качестве выхода из создавшегося положения, как правило, вводится в рассмотрение параметр времени t, который позволяет учесть динамику процесса, обусловленную с одной стороны законом сохранения внутренней среды организма, а с другой – законом развития системы.

В этом случае исходная задача преобразуется к виду:

$$Y_i = \sum_{i=1}^{k+v} \beta_i (X_i / \overline{X(t)}) + \beta_0 + u_i , \quad (i = 1,....,n) \quad (11)$$

где v – количество включенных в рассмотрение параметров внешней среды.

В данной постановке возникают новые вопросы: какую единицу времени t брать за основу – секунду, минуту, час, день; какие параметры внешней среды включать в рассмотрение, а какие – нет. К тому же сбор такого рода исходного массива данных требует дополнительных материальных и временных затрат.

На самом деле решение задачи корректировки сезонных колебаний в данной постановке можно существенно упростить ввиду следующего обстоятельства. Как известно, в математике единица времени t равнозначна, но для живых систем это далеко не так. Уровень функциональной активности организма существенным образом зависит от времени суток, года. Этот факт говорит о том, что при исследовании динамических процессов, протекающих в живых системах, следует учитывать не только количественную составляющую фактора времени, но и ее качественную компоненту. С нашей точки зрения, параметры внешней среды идеально подходят для описания качественной составляющей единицы времени. Более того, если использовать количественную составляющую временного фактора в качестве ключа вида «дата – время» и фиксировать место проведения эксперимента, то отпадает необходимость в регистрации параметров окружающей среды. Действительно, наличие такого ключа с указанием точного места проведения эксперимента позволит, обратившись в соответствующие инстанции, точно восстановить основные параметры внешней среды:

– метеорологические (физические параметры воздуха);

– экологические (химическая, радиационная загрязненность воздуха, воды и почвы);

– магнитное поле Земли (магнитные бури);

– солнечная активность (числа Вольфа, площадь, интенсивность и скорость перемещения солнечных пятен, уровень солнечной радиации);

– сейсмологические волны Земли и их характеристики;

– гравитационные волны Солнца, Луны и других планет Солнечной системы.

Предлагаемый подход позволит получить ряд преимуществ:

– решение хронобиологических, медицинских, экологических задач станет более предметным, ввиду того, что в рассмотрение будут включены факторы, являющиеся непосредственной причиной сезонных колебаний;

– база данных, созданная на основе такой структуры, не будет терять своей значимости с течением времени и приобретет такой же статус, какой присущ метеорологическим, сейсмологическим и другим базам данных:

– будет создана основа для разработки нормы для показателей различных систем организма с учетом индивидуальных особенностей организма, социальных условий жизни, параметров окружающей среды и сезонности (времени года).

Список использованной литературы:

1. Агаджанян Н.А. Человек и биосфера. М., Знание, 1987 г.
2. Бриллинджер Д. Временные ряды. Обработка данных и теория. Пер. с анг. А.В. Булинского, И.Г. Журбенко, под ред. Колмогорова А.Н. М.: Мир, 1980.
3. Бурмистров Г.А. Основы способа наименьших квадратов. М., Госгеолтехиздат, 1963.
4. Джонстон Дж. Эконометрические методы. Пер. под ред. Рывкина А.А., М., Статистика, 1980.
5. Дильман В.М. Большие биологические часы. М., Знание, 1986.
6. Зуев С.М., Кадиров Р.Х. Нелинейный обобщенный показатель тяжести заболевания // Сб. науч. труд. ОВМ АН СССР «Математическое моделирование в иммунологии и медицине» /Под ред. Асаченкова А.Л. М.,1986.– С. 100-106.
7. Крамер Г. Математические методы статистики. – М.:Мир, 1975. – 650 с.
8. Линник Ю.В. Метод наименьших квадратов и основы теории обработки наблюдений. ГИФМЛ, 1962.
9. Математическое моделирование в иммунологии и медицине. – Под ред. Асаченкова А.Л. – М.: ОВМ АН СССР, 1987.
10. Погожев И.Б. Применение математических моделей заболеваний в клинической практике. – М.: Наука, 1988.– 192 с.
11. Рао С.Р. Линейные статистические методы и их применение. М.: Наука. – 1968. – С. 550.
12. Скалько Ю.И. Об одном подходе к построению и применению обобщенных индексов тяжести заболеваний //Сб. науч. труд. ОВМ АН СССР «Математическое моделирование в иммунологии и медицине» /под ред. Асаченкова А.Л., М.:1986.
13. Молчанов А.М. Математические модели в экологии. Роль критических режимов. – В кн. Математическое моделирование в биологии, М., Наука, 1975.
14. Ягодинский В.Н. Ритм, ритм, ритм! М.: Знание, 1985.
15. Aitken A.C. On Least-squares and Linear Combination of Observations. Proc., Roval Soc., Edinburg, 1934, vol. 55.
16. Jorgenson D.W. Minimum Variance, Linear, Unbiased Seasonal Adjustment of Economic Time Series. J. Amer. Statist. Assoc., 1964, vol. 59.
17. Wise J. Regression analysis of relationships between autocorrelated time series. J. Roy. Statist. Soc., series B 18, 1956, p. 240-256.