УДК 62-50

СИСТЕМНЫЙ ПОДХОД К ПОСТРОЕНИЮ РЕГРЕССИОННОЙ МОДЕЛИ ПО ВРЕМЕННЫМ РЯДАМ

П.И. БИДЮК, И.В. БАКЛАН, В.Н. РИФА

Описывается модифицированный подход к построению математических моделей разнообразных процессов. Сформулированы понятия структуры модели, которая разрешает выбрать надлежащую структуру модели в процессе ее построения. Описанный подход был успешно испытан на ряде примеров построения эконометрических моделей.

Известны методики построения моделей типа авторегрессии со скользящим средним (АРСС), АРСС с эндогенными переменными (АРССЭ) или АРСС с интегрированным скользящим средним (АРИСС) [1–3]. Однако в представленных методиках понятие структуры модели представлено нечетко, а также недостаточно внимания уделяется определению нелинейностей модели. Предлагаемый ниже системный подход может быть использован при построении линейных моделей, а также моделей с нелинейностями относительно переменных (псевдолинейные модели). Хотя модели, нелинейные относительно параметров, здесь не рассматриваются, отдельные элементы предлагаемой методики могут быть применены также при построении моделей и такого класса.

В соответствии с предлагаемым подходом построение модели по временным рядам состоит из пяти следующих этапов:

1. Выполнить анализ процесса (процессов), для которого строится модель на основании измерений входных и выходных переменных, представленных соответствующими временными рядами.

2. Выполнить анализ имеющихся временных рядов на возможное присутствие нелинейностей с помощью ряда критериев.

3. Выбрать структуры моделей-кандидатов. Для этого необходимо выполнить следующее: вычислить и выполнить анализ корреляционной матрицы для временных рядов зависимой и независимых переменных с целью определения экзогенных переменных, которые необходимо включить в модель; вычислить автокорреляционную и частную автокорреляционную функцию для зависимой переменной с целью выбора порядка авторегрессионной части модели.

4. Выбрать метод (методы) для оценивания коэффициентов (параметров) моделей-кандидатов и оценить их параметры.

5. Выбрать лучшую (адекватную) модель из полученного на четвертом этапе множества кандидатов, используя для этой цели набор статистических параметров.

СТРУКТУРА МОДЕЛИ

Прежде чем перейти к рассмотрению конкретных этапов построения модели, рассмотрим понятие структуры математической модели, которое будет использоваться в дальнейшем. Понятие структуры модели включает в себя следующее:

1. **Порядок модели.** Это порядок дифференциального, разностного или иного уравнения, используемого для описания динамики процесса или объекта. Например, стохастическое разностное авторегрессионное (АР) уравнение второго порядка имеет вид

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \varepsilon(k).$$

То есть, порядок этого разностного уравнения определяется числом задержанных во времени значений переменной, используемых в правой части уравнения. Стохастическим оно называется потому, что в правой части присутствует случайная переменная $\varepsilon(k)$, назначение которой рассмотрим далее. Следует отметить, что введение случайной составляющей обязательно требует описания ее основных (предполагаемых или известных точно) статистических характеристик, таких как математическое ожидание, дисперсия, автокорреляционная функция и коррелированность с эндогенной переменной.

2. **Размерность модели.** Она определяется числом уравнений, используемых для описания объекта или процесса. Процесс, описываемый одним уравнением, называют одномерным или скалярным. Процесс, который описывают двумя и более уравнениями, называют многомерным. Удобным является представление в пространстве состояний. При этом размерность модели соответствует размерности вектора состояния процесса (объекта).

3. **Наличие нелинейностей и их характер**. Определить наличие нелинейностей — не всегда простая задача. Так, для механических и некоторых других систем наличие нелинейностей можно определить путем предварительного изучения законов, закономерностей и особенностей их функционирования. Например, известно, что для механических систем характерными являются нелинейности типа «люфт», «трение», билинейности, а для электрических — гистерезис.

При построении регрессионных моделей чаще всего встречаются нелинейности относительно переменных и нелинейности относительно параметров. Примером нелинейности относительно переменных может быть распространенная полиномиальная стохастическая регрессия вида

$$y(k) = a_0 + a_1 x(k) + a_2 x^2(k) + a_3 x^3(k) + \varepsilon(k).$$

Коэффициенты этого уравнения можно оценивать обычным методом наименьших квадратов (МНК) при надлежащем построении матрицы измерений [4]. Еще одним примером может быть логистическое уравнение

$$y(k) = a\, y(k-1) - a\, y^2(k-1) + \varepsilon(k),$$

которое описывает нелинейные процессы при $0 < a \le 4$ и $y(0) \in (0,1)$. В предельном случае (при $a = 4$) это уравнение описывает хаотический процесс.

Нелинейность по параметрам обусловлена наличием в модели произведений коэффициентов, например, в виде

$$y(k) = a_0 + a_1 a_2 x(k) + a_2 \exp(-bx(k)) + \varepsilon(k).$$

Коэффициенты (параметры) такой модели невозможно оценить с помощью обычного МНК, поэтому для решения этой задачи используют нелинейный МНК, метод максимального правдоподобия или другие методы нелинейного оценивания.

4. **Время запаздывания** реакции на выходе объекта по отношению к входному сигналу. Запаздывание по входу, если оно известно, достаточно легко учитывается как в непрерывных, так и в дискретных моделях. Для дискретной модели в виде разностного уравнения

$$y(k) = a_0 + a_1 y(k-1) + a_2 x(k-d) + \varepsilon(k)$$

время запаздывания d представляет собой целое число, равное количеству периодов дискретизации измерений, на которое выходной сигнал запаздывает относительно входного, т.е. $d = \mathrm{int}[\tau / T_s]$, где τ — величина запаздывания в непрерывном времени; T_s — период дискретизации измерений. Длительность периода дискретизации измерений зависит от динамики конкретного процесса и может изменяться в пределах от нескольких микросекунд для физико-технических системах до одного года в макроэкономике.

5. **Возмущения**, действующие на процесс, и способ их учета. Под возмущениями понимают входные воздействия процесса, которые оказывают, как правило, отрицательное влияние на его протекание, и не используются как управляющие. Возмущения делят на детерминированные и стохастические, а учитываются они в аддитивной или мультипликативной форме. Выше мы привели разностные уравнения, в которые возмущение $\varepsilon(k)$ входит в аддитивной форме. Пример мультипликативной формы:

$$h(k) = \upsilon(k)[\alpha_0 + \alpha_1 h(k-1)],$$

где $\upsilon(k)$ — мультипликативное возмущение. Введение случайной составляющей в модель обусловлено следующими основными причинами: присутствие неконтролируемых внешних возмущений, введение в модель излишних объясняющих переменных или, наоборот, отсутствие в модели необходимых объясняющих переменных, влияние методических и вычислительных погрешностей.

Выбор структуры модели, адекватной процессу, — задача весьма не простая и решается, как правило, итеративно. Первоначально структуру модели оценивают приблизительно на основании анализа известной информации о процессе, исследования закономерностей его протекания, анализа корреляционных функций, визуального анализа данных. При этом целесообразно выбирать несколько наиболее вероятных структур

(кандидатов). Затем определяют оценки параметров моделей-кандидатов и выбирают лучшую из них, используя соответствующие статистические характеристики моделей.

Если ни одна из моделей-кандидатов не может считаться адекватной, то необходимо исследовать на информативность экспериментальные данные, которые могут быть недостаточно информативными для оценивания модели. В таком случае может потребоваться повторный или дополнительный сбор экспериментальных данных.

АНАЛИЗ ПРОЦЕССА

На этом этапе необходимо воспользоваться всей имеющейся информацией о процессе с целью определения числа его входов и выходов; логических взаимосвязей между переменными; возможного присутствия нелинейностей и их характера; определения типа возмущений, действующих на процесс; определения присутствия запаздывания на качественном и, возможно, количественном уровнях; приблизительного определения порядка процесса. В случае исследования экономических процессов необходимо установить имеется ли влияние сезонных эффектов, присутствует ли тренд (на качественном уровне); возможно, что возникнет необходимость выдвинуть гипотезу о существовании случайного тренда; есть ли участки временных рядов с существенно различающимися уровнями колебаний (присутствие гетероскедастичности); оценить необходимость использования гипотезы относительно коинтегрированности переменных. В результате анализа процесса необходимо в общем виде постулировать структуру математической модели, которая будет использоваться в дальнейшем для описания его поведения. Например, если выдвигается гипотеза о существовании гетероскедастичности, то необходимо выбрать возможный класс моделей для ее описания. То же самое касается присутствия коинтегрированности переменных или случайного тренда.

ОПРЕДЕЛЕНИЕ НАЛИЧИЯ НЕЛИНЕЙНОСТЕЙ

Для решения этой задачи можно пользоваться различными критериями. Однако при этом необходимо знать их возможности. Покажем на простом примере, что применение линейных ковариационных функций не всегда приводит к положительным результатам. Пусть при определении структуры модели не были учтены некоторые объясняющие переменные и в результате коррелированные остатки описываются следующим уравнением:

$$\xi(k) = cu(k-1)e(k-1) + e(k), \qquad (1)$$

где $e(k)$ — белый гауссовский шум; $E[e(k)] = 0$, $E[u(k)] = 0$, $E[e(k)u(k)] = 0$, то есть, переменные $e(k)$ и $u(k)$ некоррелированны и имеют нулевое среднее; c — масштабный коэффициент. Можно показать, что нормированная автокорреляционная функция остатков и нормированная

функция взаимной корреляции между входным сигналом $u(k)$ и остатками имеют вид

$$\Phi_{\xi\xi}(\tau)=\delta(\tau),\quad \Phi_{u\xi}(\tau)=0,\quad \forall\tau\,. \tag{2}$$

Однако из уравнения (1) следует, что $\xi(k)$ — коррелированная последовательность, что будет вносить смещение в оценки параметров модели. Таким образом, в общем случае линейные корреляционные методы не позволяют определить факт присутствия нелинейных эффектов и их влияние на процесс.

Для того чтобы оценить тип связи между входом и выходом (связь линейная или нелинейная) можно воспользоваться спектральной функцией высокого порядка вида

$$X_{ij}=\frac{\left|S_\omega(\omega_i,\omega_j)\right|^2}{S_\omega(\omega_i)S_\omega(\omega_j)S_\omega(\omega_i/\omega_j)}\,, \tag{3}$$

где $S_\omega(\omega_i,\omega_j)$ — биспектральная плотность мощности; $S_\omega(\omega_i)$ — спектральная плотность мощности временного ряда. При $S_\omega(\omega_i,\omega_j)=0,$ $\forall\omega_i,\omega_j$ процесс будет линейным и третий момент входного сигнала $\mu_3=0$. Однако, если $X_{ij}=\text{const},$ то процесс линейный, но $\mu_3\neq 0$.

Такой подход к определению присутствия нелинейностей имеет два недостатка. Во-первых, оценивание спектральной плотности мощности требует применения специальной предварительной обработки сигналов в виде применения временных окон, усреднения, цифровой фильтрации и т.п. Во-вторых, он не всегда может быть использован при решении задач идентификации систем, поскольку не позволяет получить оценки параметров модели в явном виде. Кроме того, при решении этих же задач не всегда есть возможность получить измерения входного сигнала или же информативный входной сигнал определяют искусственно в виде специально генерируемых последовательностей, которые не всегда можно подавать на вход объекта вследствие особенностей его функционирования.

Что касается экономических процессов, то в этом случае, как правило, нельзя поставить эксперимент с процессом. Поэтому используют только те статистические данные, которые можно реально собрать в процессе исследования. В общем случае при идентификации систем используются тремя типами сигналов: входным, выходным и возмущением. При этом входной управляющий сигнал считают независимым от возмущения. В результате оказывается невозможным выяснение некоторых типов связей. Возможно использование также дисперсинного метода определения присутствия нелинейностей, который основан на применении следующей функции:

$$\Psi_{zu}(t_1,t_2)=E_{u(t_2)}[E_{z(t_1)}[z(t_1)\,|\,u(t_2)]-E_{z(t_1)}[z(t_1)]]^2\,, \tag{4}$$

вычисляемой с помощью достаточно сложного интегрального уравнения, если известны соответствующие плотности распределения вероятностей сигналов, которые не всегда можно определить.

В связи с вышесказанным для обнаружения нелинейностей представляется целесообразным использовать более простые корреляционные процедуры. Пусть система представлена в аналитической форме с помощью рядов Вольтерра:

$$z(t)=\sum_{k=1}^{\infty}\int_{-\infty}^{+\infty}\ldots\int h_n(\tau_1,\tau_2,\ldots,\tau_n)\,\Pi_{i=1,\ldots,n}\,u(t-\tau_i)\,d\tau+e(t)\,. \qquad (6)$$

Используя операторное представление, запишем это уравнение в виде

$$z(t)=\sum_{n=1}^{\infty}H_n[u(t)]+e(t)=H[u(t)]+e(t)=\sum_{n=1}^{\infty}H_n(u^n(t))+e(t), \qquad (7)$$

где квадратные скобки указывают на то, что H — оператор для $u(t)$, а круглые скобки — на фактическую зависимость.

В дальнейшем будем полагать, что случайные сигналы, встречающиеся в процессе идентификации, являются эргодическими, то есть средние значения по ансамблю могут быть преобразованы в средние по времени с помощью некоторой выборочной функции. Рассмотрим чувствительность модели Вольтерра второго порядка к входному сигналу $u(t)+b$. В данном случае выходной сигнал определяется так:

$$z(t)=H_1[u(t)+b]+H_2[u(t)+b]+e(t)=$$

$$=H_1(u(t)+b)+H_2(u_2(t)+2bu(t)+b^2)+e(t).$$

Если вычесть среднее с выходной величины, то получим

$$z'(t)=H_1(u(t))+H_2(u^2(t)+2bu(t)-\bar{u}^2(t))+e'(t)\,, \qquad (8)$$

где штрихом обозначен процесс с нулевым средним. Модель (8) включает зависимость от $\sigma_u{}^2=u^2(t)$ и от b, поэтому она будет давать правильный прогноз только в том случае, когда входной сигнал имеет такую же характеристику. Таким образом, чувствительность модели к входному сигналу зависит от ее типа, то есть, от ее структуры.

Для того чтобы выходной сигнал не зависел от дисперсии входного, вычтем из последнего уравнения среднее при $u(t)=0$, то есть

$$\bar{z}_b(t)=H_1[b]+H_2[b]+\ldots+e(t).$$

В результате получим следующую зависимость:

$$z'_b(t)=z(t)-\bar{z}_b(t)=H_1(u(t))+H_2(u^2(t)+2bu(t))+\ldots+e'(t)\,. \qquad (9)$$

Из (8) и (9) следует, что $\bar{z}_b(t)=\bar{z}(t)$ тогда и только тогда, когда объект линейный. Таким образом, последнее равенство можно использовать как простой тест на присутствие нелинейности.

Задачу обнаружения нелинейностей сформулируем следующим образом: требуется установить необходимость применения нелинейной модели для описания конкретной выборки данных. Для решения задачи будем пользоваться корреляционными функциями.

Пусть входной сигнал $u(t)$ и шум $e(t)$ — независимые процессы с нулевым средним и пусть все моменты с нечетными степенями для этих сигналов равны нулю, а для входного сигнала существуют все моменты с четными степенями. Рассмотрим корреляционную функцию $\Phi_{z'z'^2}(\tau)$, где $z'(t)$ — отклик системы на входной сигнал $u(t)+b$ после удаления из него среднего значения. Согласно определению, корреляционная функция

$$\Phi_{z'z'^2}(\tau) = E[z'(t+\tau)(z'(t))^2], \tag{10}$$

где

$$z'(t+\tau) = \int h_1(\tau_1)(u(t-\tau_1+\tau)+b)\,d\tau_1 +$$

$$+\iint h_2(\tau_1,\tau_2)(u(t-\tau_1+\tau)+b(u(t-\tau_2+\tau)+b)\,d\tau_1\,d\tau_2 + \ldots + e(t+\tau). \tag{11}$$

После замены переменных в уравнении (10) получим

$$z'(t+\tau) = \int h_1(t-\tau_1+\tau)u(\tau_1)\,d\tau_1 +$$

$$+\iint h_2(t-\tau_1+\tau, t-\tau_2+\tau)(u(\tau_1)u(\tau_2)+bu(\tau_1)+bu(\tau_2))\,d\tau_1\,d\tau_2 + \ldots$$

$$-\iint h_2(t-\tau_1+\tau, t-\tau_2+\tau)\overline{u}(\tau_1)\overline{u}(\tau_2)\,d\tau_1\,d\tau_2 + \ldots + e'(t+\tau). \tag{12}$$

С учетом (7) последнее уравнение запишем в виде

$$z'(t+\tau) = H_1{}^{\tau}(u(t)) + H_2{}^{\tau}(u^2(t)) + 2bH_2{}^{\tau}(u(t)) -$$

$$-H_2{}^{\tau}(\overline{u}^2(t)) + \ldots + e'(t+\tau). \tag{13}$$

Теперь функция (10) принимает вид

$$\Phi_{z'z'^2}(\tau) = E[z'(t+\tau)(z'(t))^2] =$$

$$E\left\{\begin{array}{l}\left[H_1(u)+H_2(u^2+2bu-\overline{u}^2)+H_3(u^3+3bu^2+3b^2u-3b\overline{u}^2+\ldots+e'(t)\right]^2\times\\ [H_1{}^{\tau}(u)+H_2{}^{\tau}(u^2+2bu-\overline{u}^2)+H_3{}^{\tau}(u^3+3bu^2+3b^2u-3\overline{b}\,\overline{u}^2+\ldots\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\ldots+e'(t+\tau)]\end{array}\right\} =$$

$$= E\{[(H_1H_1)(u^2)+2(H_1H_2)(u^3+2bu^2-\overline{u}\,\overline{u}^2)+2(H_1H_3)(u^4+3bu^3+3b^2u^2-$$

$$-3b\bar{u}^2u)+...+e'^2(t)]\times[H_1{}^\tau(u)+H_2{}^\tau(u^2+2bu-\bar{u}^2)+$$

$$+H_3{}^\tau(u^3+3bu^2+3b^2u-3b\bar{u}^2)+...+e'(t+\tau)]\}. \quad (14)$$

Выполним анализ корреляционной функции $\Phi_{z'z'^2}(\tau)$. Рассмотрим отдельно каждый член уравнения (14) с учетом того, что все нечетные моменты входного сигнала равны нулю, а четные — присутствуют. В результате получаем

$$E[(H_1H_1)(u^2)H_1{}^\tau(u)]=E[(H_1H_1H_1{}^\tau)(u^3)]=0. \quad (15)$$

$$E[(H_1H_1)(u^2)H_2{}^\tau(u^2+2bu-\bar{u}^2)]=$$

$$=E[(H_1H_1H_2{}^\tau)(u^4+2bu^3-\bar{u}^2u^2)]\neq 0. \quad (16)$$

$$E[(H_1H_1(u^2)H_3{}^\tau(u^3+3bu^2+3b^2u-3b\bar{u}^2)]=$$

$$=E[(H_1H_1H_3{}^\tau)(u^5+3bu^4+3b^2u^3-3b\bar{u}^2u^2)]\neq 0. \quad (17)$$

$$E[(2H_1H_2)(u^3+2bu^2-u\bar{u}^2)H_1{}^\tau(u)]=$$

$$=E[(2H_1H_2H_1{}^\tau)(u^4+2bu^3-\bar{u}^2u^2)]\neq 0. \quad (18)$$

По аналогии можно показать, что все остальные члены (за исключением тех, что содержат сигнал ошибки $e(t)$) также не равняются нулю и влияют на значение корреляционной функции. Нулевые функции имеют вид

$$E[(H_1H_1)(u^2)e'(t+\tau)]=0,$$

$$E[(2H_1H_2)(u^3+2bu^2-\bar{u}^2u)e'(t+\tau)]=0,$$

. . . . . .

$$E[(e'^2)e'(t+\tau)]=0.$$

Из приведенного анализа следует

$$\Phi_{z'z'^2}(\tau)=0, \quad \forall\tau \quad (19)$$

тогда и только тогда, когда объект линейный, то есть $H_2,H_3,...,H_n=0$. Таким образом, объект будет содержать нелинейности, когда $\Phi_{z'z'^2}(\tau)\neq 0$.

Гипотеза относительно равенства нулю третьего момента входного сигнала выполняется при возбуждении объекта равномерно распределенным гауссовским шумом и другими случайными процессами. Она проверяется с помощью следующей ковариационной функции:

$$E[u(t)u(t+\tau_1)u(t+\tau_1)], \quad \forall \tau_1, \tau_2 .$$

Присутствие во входном сигнале постоянной b способствует обнаружению нелинейностей системы, которые влияют на величину $\Phi_{z'z'^2}(\tau)$. Если положить $b=0$, то третий член разложения (16)–(19) будет равняться нулю и с помощью функции $\Phi_{z'z'^2}(\tau)$ будет невозможно определить нелинейности нечетного порядка. При наличии измерений величины $z'_b(t)$ результат, подобный (20), можно получить также для функции $\Phi_{z'_b z'^2_b}(\tau)$.

Кроме рассмотренных подходов к определению наличия нелинейностей при построении регрессионных моделей можно воспользоваться более простыми тестами. Например, статистикой [4]

$$\hat{F} = \frac{\dfrac{1}{k-2}\sum_{i=1}^{k} n_i (\bar{y}_i - \hat{y}_i)^2}{\dfrac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2},$$

где k — число групп данных; n_i — число измерений в группе; n — общее число измерений. Фактически, данная статистика представляет собой следующее отношение:

$$\hat{F} = \frac{\text{Отклонение средних значений от прямой регрессии}}{\text{Отклонение значений } y(k) \text{ от групповых средних}}.$$

Если статистика $\hat{F}$ с $\nu_1 = k-2$, $\nu_2 = n-k$ степенями свободы достигает или превосходит уровень значимости, то гипотезу о линейности нужно отбросить. Недостатком этого метода является то, что для его использования необходимо иметь не менее трех реализаций процесса, что возможно далеко не всегда.

ВЫБОР СТРУКТУРЫ МОДЕЛЕЙ-КАНДИДАТОВ

Коэффициент корреляции, а в общем случае корреляционная функция, позволяют установить наличие связи между эндогенными (зависимыми) и экзогенными (независимыми) переменными. Корреляция может быть линейная или нелинейная в зависимости от типа зависимости, фактически существующей между переменными. В большинстве практических случаев рассматривают линейную корреляцию (взаимосвязь), однако более глубокий анализ требует привлечения для исследования процессов нелинейных

зависимостей. Сложную нелинейную зависимость можно упростить, но знать о ее существовании необходимо для того, чтобы построить адекватную модель процесса.

Корреляционная матрица позволяет установить факт наличия связи между указанными переменными. Рассмотрим корреляционную матрицу размерности 3×3, которая строится для трех переменных x, y, z:

$$R=\begin{bmatrix} r_{yy} & r_{xy} & r_{zy} \\ r_{yx} & r_{xx} & r_{zx} \\ r_{yz} & r_{xz} & r_{zz} \end{bmatrix}, \tag{20}$$

где

$$r_{yx}=r_{xy},\quad r_{yz}=r_{zy},\quad r_{xz}=r_{zx}.$$

Пусть y — зависимая переменная, а x, z — технологические параметры, которые предположительно влияют на y. То есть, мы определяем наличие зависимости вида

$$y=f(x,z),$$

которая может быть представлена в форме регрессии переменной y на независимые переменные x, z:

$$y(k)=a_0+a_1x(k)+a_2z(k)+\varepsilon(k), \tag{21}$$

где k — дискретное время (например, в секундах, минутах, часах, днях, неделях, месяцах и т.д.); $\varepsilon(k)$ — случайная переменная, причины введения которой в модель следующие: наличие случайных возмущений, неучтенные регрессоры, избыточные регрессоры и ошибки вычислений.

Зачастую считают, что совокупное влияние всех указанных факторов можно с некоторым допущением описать случайной переменной $\varepsilon(k)$. Поскольку она не измеряется, то оценить ее значение (ошибку модели или остаток) можно только после оценивания коэффициентов модели, то есть

$$\varepsilon(k)\approx e(k)=\hat{y}(k)-y(k),$$

где $\hat{y}(k)$ — оценка переменной $y(k)$, полученная по модели; $y(k)$ — измерение.

Для вычисления элементов матрицы R необходимо иметь синхронные по времени выборки значений всех трех переменных y, x, z. Формула для расчета коэффициентов корреляции имеет вид

$$r_{yx}=\frac{1}{N}\frac{\sum_{k=1}^{N}\{[x(k)-\bar{x}][y(k)-\bar{y}]\}}{\sigma_x\sigma_y}. \tag{22}$$

Здесь $\bar{x}, \bar{y}$ — средние выборочные значения переменных x, y; σ_x, σ_y — стандартные отклонения этих переменных, то есть

$$\sigma_y = \sqrt{\sigma_y^2} = \left[\frac{1}{N-1} \sum_{k=1}^{N} [y(k) - \bar{y}]^2 \right]^{1/2},$$

где N — число измерений переменной y.

Коэффициенты корреляции показывают степень взаимосвязи между переменными. Очевидно, что прежде чем формально вычислять коэффициенты корреляции, необходимо выполнить анализ процесса и определить присутствие (или отсутствие) логической связи между переменными. Это позволяет ввести в рассмотрение только те переменные, которые действительно влияют на зависимую. Очевидно, что для правильного выбора переменных необходимо достаточно глубоко знать моделируемый процесс (для решения этой задачи введен первый этап).

На основании значений коэффициентов корреляции принимается решение о включении их в уравнение регрессии:

$$y(k) = a_0 + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

которое может быть представлено в общем виде как множественная регрессия:

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + \ldots + a_p x_p(k) + \varepsilon(k). \qquad (23)$$

Известно, что между коэффициентами регрессии b_1, b_2 и коэффициентами корреляции r_{yx}, r_{yz} существует однозначная взаимосвязь.

Уравнение (23) представляет собой множественную линейную регрессию p-го порядка, хотя зачастую приходится применять более сложные нелинейные модели. Характерным представителем нелинейной по переменным регрессии является полиномиальная регрессия произвольного порядка.

Для определения необходимости включения в уравнение регрессии авторегрессионной составляющей необходимо вычислить и исследовать выборочную автокорреляционную и частную автокорреляционную функцию переменной $y(k)$. Уравнение с авторегрессионной составляющей имеет вид

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + b_1 x(k) + b_2 z(k), \qquad (24)$$

то есть, в уравнение регрессии добавлена авторегрессионная (АР) составляющая второго порядка. Порядок авторегрессии определяется с помощью автокорреляционной функции. Число коэффициентов автокорреляционной функции, которые отличны от нуля в статистическом смысле, и будет составлять порядок авторегрессии.

Коэффициенты автокорреляционной функции вычисляют по формуле

$$r_y(s) = r_{y(k)y(k-s)} = \frac{1}{N} \frac{\sum_{k=s+1}^{N} \{[y(k) - \bar{y}][y(k-s) - \bar{y}]\}}{\sigma_y^2}, \quad s = 1,2,3,... , \qquad (25)$$

где σ_y^2 — выборочная дисперсия переменной $y(k)$. Число коэффициентов АКФ, отличных от нуля в статистическом смысле, указывает на порядок авторегрессионной части модели.

Уточнить порядок авторегрессионной составляющей позволяет частная автокорреляционная функция (ЧАКФ), которая вычисляется в соответствии с выражениями:

$$\Phi_{11} = r(1), \quad \Phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2}; \quad \Phi_{ss} = \frac{r_s - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_{s-j}}{1 - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_j}. \qquad (26)$$

ЧАКФ четче отражает порядок АР-модели благодаря отсутствию влияния промежуточных коэффициентов корреляции на выбранные значения переменной, то есть, коэффициент Φ_{11} характеризует степень взаимосвязи между стоящими рядом (по времени) значениями переменной, а Φ_{22} — взаимосвязь между значениями переменной, отстоящими на расстоянии двух периодов дискретизации.

Значения коэффициентов выборочной (то есть, вычисленной по выборке экспериментальных данных) частной автокорреляционной функции можно приближенно определить по экспериментальным данным следующим образом. Коэффициент a_{11} модели

$$y(k) = a_{11} y(k-1) + \varepsilon(k)$$

можно поставить в соответствие коэффициенту ЧАКФ $a_{11} \approx \Phi_{11}$, а коэффициент a_{22} модели

$$y(k) = a_{22} y(k-2) + \varepsilon(k)$$

приблизительно равняется коэффициенту Φ_{22}. Коэффициенты a_{11}, a_{22} оценивают, например, методом наименьших квадратов.

Когда мы говорим, что значения коэффициентов автокорреляционной функции должны быть отличными от нуля в статистическом смысле, это означает, что существует некоторое выражение, позволяющее установить или опровергнуть этот факт. Одним из общепринятых подходов к определению того факта, что коэффициенты АКФ существенно отличны от нуля в статистическом смысле, есть вычисление статистического параметра (или просто статистики) Льюнга-Бокса $Q(r_k)$, который вычисляется по формуле [3, 5]:

$$Q(r_k) = N(N+2) \sum_{k=1}^{s} r_k^2 / (N-k),$$

где N — длина выборки данных переменной, для которой найдены значения автокорреляционной функции r_k; s — число коэффициентов АКФ, исследуемых на существенное отличие от нуля (как правило выбирают $s \approx N/4$.

Третий этап заканчивается выбором структур нескольких моделей-кандидатов, коэффициенты которых будут оцениваться на следующем этапе.

ОЦЕНИВАНИЕ КОЭФФИЦИЕНТОВ МОДЕЛЕЙ-КАНДИДАТОВ

Вычисляем оценки коэффициентов моделей-кандидатов, которые различаются своей структурой. Например, можно выбрать авторегрессионную часть (модель) первого, второго и третьего порядков. Можно рассмотреть модели, включающие по отдельности объясняющие переменные, а также модели, содержащие все объясняющие переменные вместе. Наиболее распространенными методами оценивания параметров модели являются следующие: метод наименьших квадратов (МНК) и его модификации; метод максимального правдоподобия (ММП); метод вспомогательной переменной (МВП); нелинейный метод наименьших квадратов (НМНК) и их рекурсивные версии.

Для получения несмещенных оценок вектора параметров θ регрессионной модели с помощью метода наименьших квадратов необходимо выполнить следующие условия:

а) $\varepsilon(k)$ — некоррелированная последовательность случайных чисел с нулевым средним, то есть $E[\varepsilon(k)] = 0$, $\operatorname{cov}[\varepsilon(k)] = E[\varepsilon(k)\varepsilon(j)] = \begin{cases} \sigma_\varepsilon^2, & k = j; \\ 0, & k \neq j. \end{cases}$

б) последовательности $\varepsilon(k)$ и $y(k)$ не должны быть коррелированны между собой.

ВЫБОР ЛУЧШЕЙ МОДЕЛИ ИЗ МНОЖЕСТВА ПОЛУЧЕННЫХ КАНДИДАТОВ

Выбираем лучшую линейную или псевдолинейную модель с помощью множества статистических параметров. Они позволяют оценить по отдельности значимость коэффициентов математической модели в статистическом смысле, определить интегральную ошибку модели по отношению к исходному временному ряду, установить наличие корреляции между значениями ошибки модели (напоминаем, что они должны быть не коррелированными), а также определить степень адекватности модели физическому процессу в целом. В это множество входят следующие статистические параметры.

1. **Статистика Стьюдента**. Значимость каждого коэффициента регрессии в статистическом смысле определяют с помощью t-статистики (статистика Стьюдента), которая, как правило, вычисляется всеми пакетами статистических программ по формуле:

$$t_a = \frac{\hat{a} - a_0}{SE_a},$$

где $\hat{a}$ — оценка коэффициента, полученная с помощью пакета; a_0 — нуль-гипотеза в отношении значения этого коэффициента (обычно $a_0 = 0$); SE_a — стандартная ошибка оценки коэффициента, которая вычисляется пакетом. Очевидно, что чем меньше значение стандартной ошибки, тем лучшей является оценка коэффициента для модели.

Для определения значимости коэффициента необходимо знать длину выборки N, число оцениваемых параметров p и задаться уровнем значимости α (обычно задаются $\alpha = 1\%$, $\alpha = 5\%$ или $\alpha = 10\%$). Уровень значимости, равный 5%, означает, что при оценивании регрессии мы допускаем, что ошибочное принятие решения о значимости оценок возможно в 5% случаев. Эти параметры позволяют выбрать по таблицам значение $t_{\text{кр}}$. Если

$$-t_{\text{кр}} < t_a < t_{\text{кр}},$$

то нуль-гипотеза о незначимости коэффициента принимается; в противном случае она отвергается и коэффициент считается значимым. Поскольку значение статистики t_a обратно пропорционально стандартной ошибке SE_a, то чем большим будет значение t_a, тем более высокой будет значимость конкретного коэффициента.

2. **Коэффициент детерминации** R^2. В качестве меры информативности временного ряда часто используют его дисперсию. Коэффициент R^2 — это отношение дисперсии той части временного ряда основной переменной, которая описывается полученным уравнением, к выборочной дисперсии этой переменной,

$$R^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)}.$$

Очевидно, что для адекватной модели коэффициент детерминации должен стремиться к единице, то есть $R^2 \to 1$.

3. **Сумма квадратов ошибок модели** $\sum e^2(k)$, то есть

$$SSE = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2,$$

где, например, $\hat{y}(k) = \hat{a}_0 + \hat{a}_1 \hat{y}(k-1) + \hat{a}_2 \hat{y}(k-2) + \hat{b}_1 x(k) + \hat{b}_2 z(k)$; $y(k)$ — измерения; N — длина выборки. Очевидно, что из возможных кандидатов необходимо выбирать ту модель, для которой $\sum e^2(k)$ принимает минимальное значение.

4. **Информационный критерий Акайке (AIC).** Этот критерий учитывает сумму квадратов ошибок, число измерений N и число оцениваемых параметров p:

$$AIC = N \ln\left[\sum_{k=1}^{N} e^2(k)\right] + 2p,$$

где p — число оцененных параметров. Очевидно, что для лучшей модели критерий имеет меньшее значение, поскольку он зависит от суммы квадратов ошибок (СКО). Однако кроме СКО данный критерий учитывает длину выборки и число оцениваемых параметров, что делает его более информативным.

5. **Критерий Байеса–Шварца (BSC).** Данный критерий похож на предыдущий, однако он учитывает дополнительно длину выборки с помощью члена $\ln(N)$:

$$BSC = N \ln\left[\sum_{k=1}^{N} e^2(k)\right] + p \ln(N).$$

Его используют при длинных выборках измерительных данных**.**

6. **Статистика Дарбина-Уотсона (Durbin-Watson).** Статистика Дарбина-Уотсона вычисляется по формуле:

$$DW = 2 - 2\rho,$$

где ρ — коэффициент корреляции между значениями случайной переменной $\varepsilon(k) \approx e(k)$, $\rho = \mathrm{cov}[e(k)] = E[e(k)e(k-1)]$. Этот параметр позволяет определить степень коррелированности ошибок модели. При полном отсутствии корреляции между ошибками $DW = 2$, то есть это наиболее приемлемое значение данного параметра.

7. **Статистика Фишера** F определяет степень адекватности модели в целом. Для адекватной модели выполняется условие:

$$F > F_{\text{кр}},$$

где значение $F_{\text{кр}}$ определяется по таблице аналогично t-статистике; значение F пропорционально $R^2/(1-R^2)$, где R^2 — коэффициент детерминации. Таким образом, большему значению F соответствует более адекватная модель.

ПРИМЕР ПОСТРОЕНИЯ МОДЕЛИ

Рассмотренную выше методику проиллюстрируем при построении модели процесса на основе выборки данных из 120 измерений. Для предварительной оценки порядка авторегрессионной модели были вычислены автокорреляционная и частная автокорреляционная функции. В результате исследования АКФ и ЧАКФ установлено следующее:

1. АКФ и ЧАКФ быстро сходятся к нулевым значениям.

2. Теоретическая АКФ процесса скользящего среднего порядка q, то есть СС(q) спадает к нулю при значении запаздывания q, а теоретическая АКФ процесса АР(1) спадает к нулю геометрически. В соответствии со значениями АКФ процесс может иметь порядок 6–8, что мало соответствует действительности.

3. Коэффициенты ЧАКФ имели такие значения: $\Phi_{1,1} = 0{,}609$; $\Phi_{2,2} = 0{,}252$. В целом из анализа ЧАКФ можно сделать вывод, что порядок авторегрессии может принимать значения 1 или 2. С другой стороны, анализ АКФ свидетельствует о том, что модель может быть АР(2) или же содержать компоненты авторегрессии и скользящего среднего.

4. Небольшой выброс АКФ при значении запаздывания 4 и увеличенное значение ЧАКФ при том же значении запаздывания свидетельствуют о существовании влияния входной переменной, задержанной на 4 периода дискретизации измерений.

Из сказанного следует, что для математического описания процесса необходимо воспользоваться моделью АРСС(1,1) или АР(2). Возможно понадобиться введение времени запаздывания, равного 4. В табл. 1 приведены варианты оценивания нескольких возможных структур регрессионной модели.

Таблица 1. Варианты оценивания регрессионной модели

	$p=1,\ q=0$	$p=2\ \ q=0$	$p=1,\ q=1$	$p=1, q=1{,}4$	$p=1,\ q=2$
a_0	0,011 (4,14)	0,011 (3,31)	0,012 (2,63)	0,011 (2,76)	0,012 (2,62)
a_1	0,618 (8,54)	0,456 (5,11)	0,887 (14,9)	0,791 (9,21)	0,887 (13,2)
a_2		0,258 (2,89)			
β_1			-0,484 (-4,22)	-0,409 (-3,62)	-0,483 (-4,19)
β_2					-0,002 (-0,019)
β_4				0,315 (3,36)	
RSS	0,0156	0,0145	0,0141	0,0134	0,0141
AIC	-503,3	-506,1	-513,1	-518,2	-511,1
BSC	-497,7	-497,7	-504,7	-507,0	-499,9
$Q(12)$	23,6(0,08)	11,7(0,302)	11,7(0,301)	4,8(0,898)	11,7(0,301)
$Q(24)$	28,6(0,157)	15,6(0,833)	15,4(0,842)	9,3(0,991)	22,6(0,749)
$Q(30)$	40,1(0,082)	22,8(0,742)	22,7(0,749)	14,8(0,972)	22,6(0,749)

Примечание. В скобках указана t-статистика для оценок каждого коэффициента. При этом за нулевую гипотезу принято, что оценки равняются нулю. RSS (residual square sum) — сумма квадратов остатков (ошибок модели). $Q(n)$-статистика Льюнга–Бокса для автокорреляции n остатков оцениваемой модели. Для 122 измерений основной переменной $N/4 \approx 30$. В скобках приведен уровень значимости.

ВЫВОДЫ

1. Оценка модели АР(1) подтверждает результаты предварительного анализа. Статистика Люнга–Бокса для 12 задержанных значений остатков имеет значение 23,6, а поэтому можно отклонить нуль-гипотезу, что $Q = 0$ на уровне значимости 1%. Это свидетельствует о присутствии существенной последовательной корреляции между ошибками модели. Таким образом, модель АР(1) не может быть использована для математического описания использованного временного ряда.

2. Из табл. 1 видно, что модель АР(2) имеет лучшие статистические характеристики по сравнению с моделью АР(1). Оценки коэффициентов модели ($\hat{a}_1 = 0{,}456$, $\hat{a}_2 = 0{,}258$) существенно отличаются от нуля на уровне 1%, а корни характеристического уравнения находятся внутри окружности единичного радиуса. Значение Q-статистики свидетельствует о том, что автокорреляция между ошибками является статистически несущественной, то есть, нуль-гипотеза $Q = 0$ подтверждается. Критерий AIC имеет меньшее значение для модели АР(2). В целом модель АР(2) лучше аппроксимирует ряд чем АР(1).

3. Модель АРКС(1,1) имеет лучшие статистические показатели чем АР(2). Значение t-статистики для оценок коэффициентов (14,9 и -4,22) свидетельствуют о высоком качестве оценок. Оценка $\hat{a}_1 = 0{,}887$ положительная и блика к единице, а Q-статистика свидетельствует, что автокорреляция остатков не имеет статистической значимости. Критерии AIC и BSC также показывают более высокое качество модели АРКС(1,1) по сравнению с АР(2).

4. Для того чтобы выявить присутствие запаздывания на 4 периода дискретизации, в пробную модель скользящего среднего введен дополнительный член с задержкой 4. То есть, пробная модель имела вид

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k) + \beta_1 \varepsilon(k-1) + \beta_4 \varepsilon(k-4).$$

Отметим, что именно член $\beta_4 \varepsilon(k-4)$ лучше описывает эффект запаздывания (при его наличии) чем авторегрессионный член $a_4 y(k-4)$. Член скользящего среднего точнее описывает такие эффекты чем авторегрессионный. Все коэффициенты модели АРКС(1,(1,4)) имеют значительную статистическую значимость с t-статистиками, равными 9,21; -3,62 и 3,36 соответственно. Все значения Q-статистики весьма незначительны, что свидетельствует о том, что автокорреляция остатков статистически близка нулю. Критерии AIC и BSC также поддерживают преимущество модели АРКС(1, (1, 4)).

5. Для коэффициента $\hat{\beta}_2$ в последней рассмотренной пробной модели АРСС(1, 2) t-статистика имеет достаточно низкое значение, что дает основания для исключения этой модели из дальнейшего рассмотрения.

Следующим шагом исследования данного процесса может быть тестирование временного ряда на гетероскедастичность, то есть, определение стационарности дисперсии ряда.

ЛИТЕРАТУРА

1. *Бокс Дж., Дженкинс Г.* Анализ временных рядов. Т. 1, 2. — М.: 1974. — 406 с.
2. *Ивахненко А.Г., Степашко В.С.* Помехоустойчивость моделирования. — К.: Наук. думка, 1984. — 295 с.
3. *Enders W.* Applied econometric time series. — New York: Wiley and Sons, 1994. — 433 p.
4. *Закс Б.* Статистическое оценивание. — М.: Статистика, 1976. — 598 с.
5. *Бідюк П.І., Половцев О.В.* Аналіз та моделювання економічних процесів перехідного періоду. — К: ПЛАБ-75, 1999. — 230 с.

Поступила 16.08.2002