Бидюк П.И., Зворыгина Т.Ф.

СТРУКТУРНЫЙ АНАЛИЗ МЕТОДИК ПОСТРОЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ ПО ВРЕМЕННЫМ РЯДАМ НАБЛЮДЕНИЙ

В работе выполняется анализ методик построения моделей типа авторегрессии со скользящим средним (АРСС), АРСС с эндогенными переменными (АРССЭ) или АРСС с интегрированным скользящим средним (АРИСС) [1-4]. При этом указаны основные этапы моделирования временных рядов, рассмотрено понятие структуры модели и описаны варианты обычно применяемых методов и критериев на каждом из этапов.

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

- выполнить предварительный анализ имеющихся данных (на присутствие выбросов или пропусков) и априорной информации о процессах, для которых строится модель, определить цель построения модели;
- выполнить проверку имеющихся временных рядов на возможное присутствие нелинейностей;
- выбрать класс структур моделей-кандидатов, для чего необходимо: 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;
- выбрать способ генерации структур моделей в зависимости от ограничений на время решения, количества входных переменных и уровня требований к модели;
- выбрать метод (методы) оценивания коэффициентов (параметров) моделей-кандидатов и оценить их параметры;
- выбрать критерий отбора (селекции) лучшей из моделей-кандидатов;
- проверить адекватность полученной модели в целом.

Понятие структуры модели включает в себя:

- порядок модели по выходу,
- размерность выходного вектора модели,
- наличие нелинейностей и их характер,
- запаздывания реакции на выходе объекта по отношению к входному сигналу (лаговые эффекты),
- тип возмущений, действующих на процесс, и способ их учета.

Введение случайной составляющей в модель обуславливается следующими основными причинами: присутствие неконтролируемых внешних возмущений, введение в модель излишних 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 или с применением некоторого метода регулярного перебора вариантов. 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; выяснения логических взаимосвязей между переменными; установления возможного присутствия нелинейностей и их характера; определения типа возмущений, действующих на процесс; определения присутствия запаздываний 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, возникнет необходимость выдвинуть гипотезу о существовании случайного тренда; есть ли участки временных рядов с существенно различающимися уровнями колебаний (присутствие гетероскедастичности); оценить необходимость использования гипотезы о 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 об их возможностях.

При построении регрессионных моделей можно воспользоваться простыми тестами, например, статистикой [5]

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

где $k-$ число групп данных; n_i- число измерений в группе; $n-$ общее число измерений. Фактически эта статистика представляет собой отношение отклонений средних значений от прямой регрессии к отклонениям значений $y(k)$ от групповых средних. Если статистика $\widehat{F}$ с $\nu_1 = k-2, \nu_2 = n-k$ степенями свободы превосходит уровень значимости, то гипотезу о линейности нужно отбросить.

При этом нужно помнить, что этой статистикой можно обоснованно пользоваться лишь в случае, когда структура модели задана. Если же существуют несколько возможных структур, то возникнут сложности, связанные с тем, что в статистику входят оценки $\hat{y}_i$.

Выбор класса структур 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.

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).

Для определения необходимости включения в уравнение регрессии авторегрессионной составляющей следует вычислить и исследовать выборочную

автокорреляционную и частную автокорреляционную функцию переменной $y(k)$. Уравнение с авторегрессионной составляющей имеет вид:

$$y(k) = a_0 + \sum_i a_i x_i(k) + \sum_j b_j y(k-j+1), \qquad (2)$$

то есть в уравнение регрессии добавлена авторегрессионная (АР) составляющая. 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 вычисляются по формуле:

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

где $s_y^2 -$ 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:

$$\Phi_{11} = r(1), \qquad \Phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2}; \quad \ldots, \quad \Phi_{ll} = \frac{r_l - \sum_{j=1}^{l-1} \Phi_{l-1,j} r_{l-j}}{1 - \sum_{j=1}^{l-1} \Phi_{l-1,j} r_j}. \qquad (4)$$

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} характеризует взаимосвязь между значениями переменной, отстоящими на расстоянии двух периодов дискретизации.

Когда говорят, что значения коэффициентов автокорреляционной функции должны быть отличными от нуля в статистическом смысле, это означает, что существует некоторое выражение, которое позволяет подтвердить или опровергнуть этот факт. Одним из общепринятых подходов к определению того, что коэффициенты АКФ существенно отличны от нуля в статистическом смысле, есть вычисление статистического параметра (статистики) Льюнга-Бокса $Q(r_k)$ по формуле [2, 4]:

$$Q(r_k) = N(N+2) \sum_{k=1}^{s} r_k^2 / (N-k), \qquad (5)$$

где $N-$ длина выборки данных переменной, для которой найдены значения автокорреляционной функции r_k; $s-$ число коэффициентов АКФ, которые исследуются на существенное отличие от нуля.

Более сложные процедуры выбора множества классов моделей описаны в [6].

Выбор способа генерации структур моделей

После определения порядка АР-модели можно воспользоваться генератором структур моделей различной сложности с применением некоторого метода регулярного перебора вариантов при заданном числе запаздываний. Для выбора наиболее подходящего генератора структур необходимо учесть такие факторы, как количество входных переменных, ограничения на время решения и уровень требований к модели.

Заметим, что при ограниченном времени и простой модели применяются методы включения и исключения, при необходимости получить более достоверную модель – метод включения-исключения [7], а в случае большого количества переменных применяются методы ветвей и границ, а также различные методы из семейства МГУА [8] – комбинаторно-селекционный, линейный или нелинейный многорядный МГУА.

Выбор критерия селекции моделей

На этом этапе выбирают лучшую линейную или псевдолинейную (линейную по коэффициентам) модель из множества моделей-претендентов. Критерий селекции моделей зависит от типа возмущений, влияющих на процесс, и целей, которые преследуются при моделировании.

При вычислении критериев часто используется *остаточная сумма квадратов ошибок модели*:

$$RSS = \sum_{k=1}^{n} [\hat{y}(k) - y(k)]^2, \qquad (6)$$

где $\hat{y}(k)$ - оценки, $y(k)-$ измерения; $n-$ длина выборки

Сама эта величина не может служить критерием для выбора структуры, поскольку при увеличении сложности модели s происходит все более точное приближение к входным данным, что допустимо только при отсутствии возмущений.

Если известно, что шум распределен по нормальному закону, то применяются следующие критерии:

Скорректированный RSS: RSS/(n-s)

Статистика Фишера:

$$F(s) = \frac{n}{n-s} \frac{RSS(s)}{\left\| y - \hat{y} \right\|^2} \qquad (7)$$

Если возможно получить оценку дисперсии шума**,** применяется *критерий Маллоуза*

$$C_s = \frac{1}{\hat{\sigma}^2} RSS(s) + 2s - n \qquad (8)$$

Единственный критерий, который применим при любом известном распределении шума – это *информационный критерий Акаике* (AIC)**.**

$$AIC(s) = -2\ln f(x;\hat{\theta}(s)) + 2s \qquad (9)$$

Этот критерий существенно ограничивает рост сложности модели наличием аддитивного члена $2s$. Однако проблема его применения состоит в том, что в практических задачах функция распределения шума неизвестна.

В частном случае нормального шума он принимает вид критерия Маллоуза. При этом на практике он применяется в виде

$$AIC(s) = RSS(s) + 2s \qquad (10)$$

В последнее время эта формула называется критерием Акаике-Маллоуза [9].

Популярным является критерий, называемый *«финальная ошибка прогнозирования»*, который не требует дополнительной информации и вычисляется так:

$$FPE(s) = \frac{n+s}{n-s} RSS(s) \qquad (11)$$

Критерии с разбиением выборки

Если статистические оценки данных не известны, то применяются внешние критерии с разбиением выборки («перекрестного обоснования») [8].

Здесь рассмотрим разбиение на три непересекающихся подвыборки (подмножества точек) A, B, C, причем обозначим также $A \cup B = W$. Для стандартизации записи формул примем следующие обозначения: оценка параметров по МНК на некоторой подвыборке G равна

$$\theta_G = (X_G^T X_G)^{-1} X_G^T y_G, G = A, B, W, \qquad (12)$$

а значение ошибки на некоторой подвыборке Q по модели, оценки параметров которой вычислены на G , равно

$$\Delta(Q \mid G) = \left\| y_Q - X_Q \hat{\theta}_G \right\|^2, \qquad (13)$$

где Q = A, B, W, C. Разбиение на три подвыборки соответствует следующему:

$$X = \begin{bmatrix} X_A \\ X_B \\ X_C \end{bmatrix} = \begin{bmatrix} X_W \\ X_C \end{bmatrix}, \qquad Y = \begin{bmatrix} Y_A \\ Y_B \\ Y_C \end{bmatrix} = \begin{bmatrix} Y_W \\ Y_C \end{bmatrix}. \qquad (14)$$

Очевидно, что при Q = G величина $\varepsilon = \Delta(G|G)$ является остаточной суммой квадратов RSS, а при $G \neq Q$ она является *критерием регулярности*:

$$AR_B = \Delta(B \mid A), AR_A = \Delta(A \mid B). \qquad (15)$$

Тогда критерий *симметричной регулярности* имеет вид:

$$AD = \Delta(B \mid A) + \Delta(A \mid B). \qquad (16)$$

Критерий *непротиворечивости (несмещенности)* определяется выражением:

$$CB = \left\| X_W \hat{\theta}_A - X_W \hat{\theta}_B \right\|^2 = (\hat{\theta}_A - \hat{\theta}_B)^T X_W^T X_W (\hat{\theta}_A - \hat{\theta}_B) = \\ = AS - 2(\varepsilon_A + \varepsilon_B). \qquad (17)$$

Широко известен *критерий Кейна,* применяемый только тогда, когда известно, что шум нормальный:

$$K(s) = \frac{\frac{1}{n-2s}[RSS_W(s) - RSS_A(s) + RSS_B(s)]}{\frac{1}{s}[RSS_A(s) + RSS_B(s)]} \qquad (18)$$

Существует еще один популярный критерий, называемый *критерий «скользящего контроля», «усредненный критерий регулярности», или «джек-найф»:*

$$JN(s) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f_i(x_i; \hat{\theta}_i(s))]^2 \qquad (19)$$

Простого перечня возможных критериев не достаточно, так как каждому исследователю, не являющемуся экспертом в области моделирования, на практике нужно решать задачу выбора наиболее подходящего критерия из списка возможных. Для решения этой задачи необходимо провести анализ применимости (ограничений на применение) различных критериев.

Примером методики такого анализа может служить следующая таблица.

Априорная информация	**Цель моделирования**	
	Аппроксимация Интерполяция Модель вход-выход	Прогнозирование Экстраполяция
Известна функция распределения шума	Акаике	Акаике
Шум нормальный	RSS/(n-s), Фишера, Кейна, Акаике, Маллоуза	регулярности, джек-найф, Акаике
Статистические характеристики неизвестны	Регулярности, несмещенности, джек-найф,	Регулярности, джек-найф

Содержание данной конкретной таблицы не претендует на полноту и однозначность, поскольку ее заполнение зависит от мнений различных экспертов в области моделирования, обладающих своими привычками, опытом и предпочтениями. Однако составление такого рода таблиц необходимо, так как является одним из основных этапов разработки правил принятия решений в области моделирования по данным наблюдений [11].

Если говорить конкретно о критериях селекции моделей, то следующим шагом процесса выбора должен стать анализ их применимости в зависимости от количества данных, по которым строится модель. Некоторые критерии не могут работать с малым количеством данных, в то время как при достаточном их количестве может не возникнуть необходимости применять сложные критерии, можно обойтись более простыми.

Оценивание коэффициентов моделей-кандидатов

На этом этапе вычисляют оценки коэффициентов моделей-кандидатов, которые различаются своей структурой. Например, моделью-претендентом может быть авторегрессионная составляющая первого, второго и третьего порядка. Могут проверяться модели, включающие по отдельности объясняющие переменные, а также модели, которые содержат 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.

Все эти методы имеют свои условия применения. Например, для получения несмещенных оценок вектора параметров θ регрессионной модели с помощью метода наименьших квадратов необходимо проверить выполнение известных предположений классического регрессионного анализа [7].

Проверка адекватности модели.

На этом этапе оценивают степень адекватности модели природе процесса в целом и имеющимся априорным предположениям.

В принципе значения упомянутых выше критериев, применяемых для селекции моделей, являются также некоторыми характеристиками адекватности модели. Однако на практике принято использовать дополнительные оценки адекватности, такие как:

1. *t-статистика* Стьюдента. Значимость каждого из коэффициентов регрессии в статистическом смысле определяют с помощью $t-$ *статистики*, которая вычисляется по формуле [7]:

$$t_a = \frac{|\hat{a} - a_0|}{\sqrt{z_{ii}}c}, \qquad (20)$$

где $\hat{a}$ - оценка коэффициента; $a_0 -$ нуль-гипотеза в отношении значения этого коэффициента (обычно $a_0 = 0$); с – оценка дисперсии случайного возмущения , z_{ii} - i-й диагональный элемент матрицы $(X^T X)^{-1}$.

Для определения значимости коэффициента необходимо учитывать длину выборки N, число оцениваемых параметров p и задаться уровнем значимости α (обычно задаются $\alpha = 1\%, \alpha = 5\%$ или $\alpha = 10\%$). Уровень значимости указывает долю ошибочно принятых решений о значимости параметров при оценивании регрессии. Если вычисленное значение по сравнению с табличным $t_{крит}$ удовлетворяет условию

$$-t_{крит} < t_a < t_{крит}, \qquad (21)$$

то нуль-гипотеза о не значимости коэффициента принимается; в противном случае она отвергается и коэффициент считается значимым. 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. Он вычисляется по формуле:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (22)$$

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

3. *Критерий Байеса-Шварца* (BSC)**.** Данный критерий похож на критерий Акаике (9), однако он учитывает дополнительно длину выборки с помощью члена $\ln(N)$:

$$BSC = \ln\left[\sum_{k=1}^{N} e^2(k)\right] + s\ln(N), \qquad (23)$$

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

4. *Статистика Дарбина-Уотсона* (Durbin-Watson)

Статистика Дарбина-Уотсона вычисляется по формуле:

$$DW = 2 - 2\rho, \qquad (24)$$

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

На этом этапе также можно применять более сложные критерии, такие как критерий Уиттла, Хеннана, Бартлетта и тест на сериальную независимость [4].

В последние годы важнейшей дополнительной характеристикой адекватности модели считается ее проверка на дополнительной части выборки, которая не использовалась при составлении модели [2, 12]. Необходимость такого разделения выборки данных обусловлена тем, что модель с наибольшим количеством переменных, как правило, не дает лучшее качество прогноза, чем модель с оптимальным количеством переменных, хотя последняя имеет худшие аппроксимирующие свойства.

Отметим, что этот критерий (в виде проверки на третьей, *экзаменационной* части выборки данных) используется в алгоритмах МГУА с момента его зарождения, однако широкое распространение он приобрел лишь в последние годы [2].

Заключение

Рассмотренные в обзоре методы и критерии, использующиеся при моделировании по данным наблюдений, далеко не исчерпывают то множество способов, которое практически используется исследователями. Эта работа содержит общий анализ основных этапов процесса решения задачи моделирования и описание множеств решений на каждом из них. Следующим шагом работы должен стать анализ эффективности каждого из возможных решений на некоторых типовых классах прикладных задач. Его можно провести как путем анализа литературы и изучения мнений экспертов, так и путем вычислительных экспериментов. Типовые задачи определяются объемом и характером априорной информации об объекте и неопределенностях. Пример такого анализа дан в работе в виде таблицы применения критериев селекции моделей.

Литература

1. Бокс Дж., Дженкинс Г. Анализ временных рядов (т.1,2). – М.: 1974. – 406 с.
2. Кашьяп Р.Л., Рао А.Р. Построение динамических стохастических моделей по экспериментальным данным / Пер. с англ. – М: Наука, 1983. – 384 с.
3. Конева Е.С. Выбор моделей для реальных временных рядов // Автоматика и телемеханика, № 6, 1988, стр. 3-18.
4. Enders W. Applied econometric time series. – New York: Wiley & Sons, 1994. – 433 p.
5. Закс Б. Статистическое оценивание. – М.: Статистика, 1976. – 598 с.
6. Бідюк П.І., Половцев О.В. Аналіз та моделювання економічних процесів перехідного періоду. – Київ: ПЛАБ-75, 1999. – 230 с.
7. Вучков И., Бояджиева Л, Солаков Е. Прикладной регрессионный анализ / Пер. с болг. – М: Финансы и статистика, 1987. - 239 с
8. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. – Киев.: Наукова Думка, 1984. – 295 с.
9. Стадник М.П. Модификация критерия Мэллоуза-Акаике для подбора порядка регрессионной модели / Автоматика и телемеханика. - 1988. - № 4. - С. 98-108.
10. Степашко В.С. Алгоритмы МГУА как основа автоматизации процесса моделирования по экспериментальным данным // Автоматика. - 1988. - № 4. - С.44-55.
11. Степашко В.С., Зворыгина Т.Ф. О проектировании диалоговой оболочки СППР для моделирования по данным наблюдений // Моделирование и управление состоянием эколого-экономических систем региона. - Киев, 2001. - С. 64-69.
12. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа и пакет ППСА. – Москва, Финансы и статистика, 1986 – 232 с.

Бидюк П.И., Зворыгина Т.Ф.

СТРУКТУРНЫЙ АНАЛИЗ МЕТОДИК ПОСТРОЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ ПО ВРЕМЕННЫМ РЯДАМ НАБЛЮДЕНИЙ

Задача построения авторегрессионных моделей по временным рядам наблюдений рассматривается как последовательность этапов анализа данных. Приведено краткое описание множеств возможных процедур на каждом из таких этапов с рекомендациями и ограничениями на их применение. Для одного из этапов – выбора критерия селекции моделей – приведен пример анализа применимости.

Бідюк П.И., Зворигіна Т.Ф.

СТРУКТУРНИЙ АНАЛІЗ МЕТОДИК ПОБУДОВИ РЕГРЕСІЙНИХ МОДЕЛЕЙ ЗА ЧАСОВИМИ РЯДАМИ СПОСТЕРЕЖЕНЬ

Задача побудови авторегресійних моделей за часовими рядами спостережень розглядається як послідовність етапів аналізу даних. Наведено короткий опис множин можливих процедур на кожному з таких етапів з рекомендаціями й обмеженнями на їх застосування. Для одного з етапів – вибору критерію селекції моделей – наведено приклад аналізу застосовності.

Bidiuk P.I., Zvorygina T.F.

STRUCTURAL ANALYSIS OF METHODS OF REGRESSION MODEL CONSTRUCTION AFTER TIME SERIES OF OBSERVATIONS

A problem of construction of auto regression models after the observed time series is investigated as a sequence of stages of data analysis. Short description of sets of possible procedures for each stage is provided with suggestions and restrictions concerning its application. Applicability analysis is given for the stage of definition of a model selection criterion.