УДК 338.27

Павлючук Ю.Н., д-р техн. наук, проф.
Шевчук А.А., асп.
(Брестский государственный технический университет, Республика Беларусь)

МАТЕМАТИЧЕСКИЙ АППАРАТ ПОСТРОЕНИЯ ПРОГНОЗНЫХ МОДЕЛЕЙ: АНАЛИЗ ИСХОДНЫХ ДАННЫХ

Современная экономическая действительность требует от руководителей предприятий принятия решений в кратчайшие сроки. Запоздалая реакция на изменившуюся ситуацию на рынке может привести к кризисным ситуациям на предприятии. С целью предотвращения таких ситуаций и для принятия верного решения принято использовать различные прогнозные модели.

По нашему мнению, качество используемой или вновь строящейся модели зависит от нескольких факторов: исходных данных модели, ее спецификации, критерия выбора переменных, способа оценки параметров модели и так называемого «разрешающего правила», на основании которого принимается то или иное решение с учетом результатов модели.

Очевидно, что все перечисленные факторы, за исключением первого, зависят от конкретной решаемой задачи. Поэтому в данной статье мы бы хотели более подробно остановиться на анализе исходных данных, поскольку он является общим для всех прогнозных моделей. Более того, ошибки, допущенные в начальной выборке, оказывают существенное влияние на конечный результат и могут привести к непредсказуемым последствиям. Иначе говоря, любые искажения данных приводят к смещению параметров модели и, соответственно, к снижению достоверности прогнозных оценок. Поэтому возникает необходимость в применении специальных процедур, исключающих или хотя бы снижающих нежелательные эффекты.

При статистическом анализе выборочных совокупностей предусмотрены процедуры проверки однородности данных. Формально они применимы к любой выборочной совокупности. Наблюдение, признанное неоднородным, обычно рекомендуется

удалить из выборочной совокупности. Однако выполнение подобной рекомендации остается практически безболезненным только для пространственных выборок. В случае же временных рядов, на основе которых обычно строятся прогнозные модели, возможность применения подобной процедуры полностью исключается, но проблема неоднородности остается и требует своего решения. Поэтому ниже предлагаются методы, применение которых корректно для временных рядов. Их особенность в том, что они одновременно с формальным подходом предусматривают содержательный анализ неадекватной ситуации. В некоторых случаях под сомнение могут попадать даже те данные, которые в рамках формального подхода не тестируются как неоднородные.

Общая схема реализации предлагаемого здесь подхода предусматривает два этапа: идентификацию и удаление сомнительного наблюдения; восстановление удаленного значения с помощью специальных методов.

Реализация первого этапа предусматривает возможность применения как формальных, так и неформальных процедур. Как правило, реальные действия опираются на обе процедуры с некоторым доминированием субъективной точки зрения. Любые сомнения специалистов по поводу отдельных периодов функционирования предприятия должны инициировать проведение специального анализа, результаты которого подтверждают или отвергают сомнения.

В основу формального подхода положена проверка принадлежности фактических значений доверительному интервалу, расчет которого предусмотрен многими статистическими пакетами. Если построенная модель адекватна и фактическое значение оказалось за рамками доверительного интервала, то это является серьезным основанием для того, чтобы считать данное наблюдение нехарактерным для исследуемой выборочной совокупности.

Таким образом, первый этап завершается либо списком наблюдений, которые считаются искаженными и требуют своего уточнения, либо отсутствием такового. Если список не пуст, то следующий этап реализуется в полном объеме, если пуст, то вы-

полняется только та часть второго этапа, которая касается прогнозных расчетов.

Процедуры второго этапа, к описанию которых мы переходим, являются универсальными и их можно объединить в класс методов восстановления пропусков в данных, предназначенных для построения прогнозных моделей.

Как правило, эти данные представляются в виде прямоугольных таблиц, строки которых принято называть наблюдениями, а столбцы – переменными. Для их формирования используются разные источники: статистические справочники, интернет-ресурсы, электронные базы данных, отчеты предприятий и др. К сожалению, не всегда данные, получаемые из этих источников, обладают достаточной надежностью. Более того, с их помощью не всегда удается полностью заполнить таблицы данных, т.е. некоторые наблюдения, попавшие в таблицы, оказываются некомплектными. О данных, в которых имеются некомплектные наблюдения, принято говорить как о данных с пропусками. Их использование для прогнозных расчетов требует специальных подходов, которые ориентированы в основном на применение процедур восстановления пропусков. Рассмотрим некоторые из них.

Метод исключения некомплектных наблюдений. Это самый простой прием, в соответствии с которым рекомендуются те наблюдения, в которых отсутствуют значения одной или нескольких переменных, исключать из анализа и обрабатывать только комплектные наблюдения. Такой метод легко реализуется и может быть удовлетворительным при малом числе пропусков и достаточно большом числе наблюдений. Однако в некоторых ситуациях удаление некомплектных наблюдений приводит к серьезным смещениям в получаемых оценках и поэтому является малоэффективным. Более того, для временных рядов, как было уже отмечено, способ, основанный на удалении, неприемлем.

Методы заполнения (восстановления) пропусков. Смысл восстановления в том, что после заполнения пропусков появляется возможность обрабатывать данные с помощью обычных методов. Естественно, надежность получаемых результатов в значительной степени зависит от корректности и эффективности методов, с помощью которых осуществлялось восстановление. В

практике прогнозных расчетов для восстановления пропусков используются различные подходы. Самый простой из них – *метод средних,* основанный на вычислении средних по неполным данным и заполнении этими средними пропусков. Замена средними зачастую является первым этапом более сложной, итерационной, процедуры восстановления пропусков, хотя не исключены случаи, когда такая замена может оказаться достаточно успешной.

Интересным является *метод заполнения с «пристрастным» подбором,* когда вместо пропущенных подставляются значения переменных других наблюдений. Проблема реализации этого метода в том, чтобы подобрать комплектное наблюдение в наибольшей степени близкое (похожее) к некомплектному. По сути процедура подбора такого наблюдения является некоторым упрощенным аналогом распознающей системы. Формальный подход к решению этой задачи предусматривает введение меры, с помощью которой можно оценить степень близости между парой любых наблюдений. Чаще всего для этих целей используется метрика Евклида

$$\rho(x_i, x_k) = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{kj})^2}.$$

Если требуется в i-ом наблюдении восстановить значение l-го показателя, то последовательность действий по восстановлению выглядит следующим образом:

1) рассчитывается расстояние между i-ым наблюдением с отсутствующим значением l-го показателя и всеми комплектными наблюдениями

$$\rho(x_i, x_k) = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{kj})^2}, \quad k = 1...n, k \neq i;$$

2) среди комплектных наблюдений выбирается наблюдение с номером k^*, для которого

$$\rho(x_i, x_{k^*}) = \min_k \rho(x_i, x_k);$$

3) значение l-го показателя в i-ом наблюдении заменяется на значение соответствующего показателя с номером k^*, т.е.

$$x_{il} = x_{k \cdot l}.$$

Аналогично можно восстанавливать и те некомплектные наблюдения, в которых более одного пропуска.

В тех случаях, когда переменные представляют собой динамические ряды, по преимуществу используемые в экономическом прогнозировании, и когда можно установить закономерность их изменения во времени, то для восстановления пропусков в таких переменных можно использовать *метод интерполяционных расчетов* (для этих целей используются трендовые модели). Например, если рассматриваемая ситуация выглядит таким образом, что в j-ой переменной пропущено наблюдение с номером l, то для построения трендовой модели формируется набор данных в следующем виде:

$$\begin{array}{c} x_{1j},\ x_{2j},\ \ldots,\ x_{l-1j},\ -,\ x_{l+1j},\ \ldots\ x_{nj} \\ 1,\ \ 2,\ \ldots,\ l\text{-}1,\ \ l,\ \ l\text{+}1,\ \ldots,\ \ n\ . \end{array}$$

Трендовая модель $f(t)$ строится обычным образом с использованием метода наименьших квадратов. Расчетное значение $x_{lj}^* = f_j(l)$ этой модели используется в качестве пропущенного.

Комбинированные методы восстановления пропусков. Для повышения точности имеет смысл использовать сразу несколько методов, с помощью которых восстанавливается одно и то же пропущенное значение. При комбинированном подходе каждому методу отводится своя роль. Например, с помощью одного устанавливается начальное приближение восстанавливаемых значений, а с помощью другого – осуществляется их корректировка и уточнение. Хорошей иллюстрацией такого комбинирования является подход, состоящий в последовательном применении метода средних и метода скользящего среднего. В рамках этого подхода на первом этапе среднее значение принимается за восстанавливаемое

$$\bar{x}_j = \frac{1}{n\text{-}1} \sum_{i \neq j} x_{ij},$$

а на втором – среднее заменяется скользящим средним

$$\tilde{x}_{lj} = \frac{x_{l-1\,j} + \tilde{x}_j + x_{l+1\,j}}{3}.$$

Недостаток такого подхода в том, что с его помощью восстанавливаются не все пропуски, он применим только для наблюдений, номера которых удовлетворяют неравенству $2 < l < n - 1$. Кроме того, не удается оценить точность, с которой проведено восстановление.

Более сложный вариант комбинированной процедуры получается, когда среднее, принятое за пропущенное значение на первом этапе, уточняется итерационно, путем многократного построения трендовой модели и замене предыдущей (k–1)-ой оценки пропуска текущим k-м расчетным значением $\tilde{x}_{lj}^{k} = f(t)$. Процесс уточнения восстанавливаемого значения продолжается до тех пор, пока не выполнится условие $|\tilde{x}_{lj}^{k} - \tilde{x}_{lj}^{k-1}| < \varepsilon$, где ε – некоторая заданная, достаточно малая, положительная величина.

Особый интерес представляет *подход, основанный на идее адаптивных ожиданий.* Суть его в том, что за значение восстанавливаемого пропуска принимается величина, определяемая в соответствии с принципом адаптивного ожидания по формуле

$$\tilde{x}_{lj} = \alpha x_{l-1j} + (1 - \alpha) x_{l+1j}, \quad 0 < \alpha < 1,$$

где a – параметр, который подлежит определению.

Для определения параметра α в случае восстановления единичного пропуска в монотонной последовательности данных, по крайней мере, можно предложить два подхода. Первый предусматривает получение α в виде среднего

$$\alpha^* = \frac{1}{n-5} \sum_{i \neq l} \alpha_i,$$

где a_i – параметр, определяемый по формуле

$$\alpha_i = \frac{x_{i+1\,j} - x_{ij}}{x_{i+1\,j} - x_{i-1\,j}}, \quad i = 2, \ldots, l-2, l+2, \ldots, n-1,$$

которая легко выводится из соотношения

$$x_{ij} = \alpha_i x_{i-1\,j} + (1 - \alpha_i) x_{i+1\,j}.$$

Недостаток этого подхода в том, что даже при восстановлении единственного пропуска, не используется пять наблюдений (включая сам пропуск).

В основе второго подхода лежит идея оптимального восстановления пропущенного значения в том смысле, что восстановленное значение должно быть таким, чтобы минимизировать ошибку интерполяции исходного ряда с помощью трендовой или регрессионной модели. Другими словами, параметр α должен настраиваться по критерию суммы квадратов отклонений расчетных значений от фактических.

Для реализации этого подхода сформируем набор данных следующим образом:

$$\begin{array}{cccccccc} x_{1j}, & x_{2j}, & \ldots, & x_{l-1j}, & \tilde{x}_{lj}, & x_{l+1j}, & & x_{nj} \\ 1, & 2, & \ldots, & l-1, & l, & l+1, & \ldots, & n, \end{array}$$

где $\tilde{x}_{li}$ определяется по вышеприведенной формуле.

Тогда трендовая модель, построенная по этому набору данных, окажется зависящей от настраиваемого параметра α, т.е. $f(t, \alpha)$.

Техника восстановления пропущенного значения в этом случае реализуется через процедуру, состоящую в последовательном изменении α с шагом h и построении для каждого α трендовой модели. Из всех построенных моделей для получения пропущенного значения используется α^* из той, которая дает наименьшую сумму квадратов ошибок, т.е.

$$\tilde{x}_{lj} = f(l, \alpha^*);$$

$$\alpha^* = Arg \min_{\alpha} \sum_{t} (x_{tj} - f(t, \alpha))^2.$$

Рассмотрим случай, когда возникает необходимость в восстановлении сразу двух пропусков, которые в исходных данных расположены рядом, т.е. структура данных имеет следующий вид:

$$\begin{array}{cccccccccc} x_{1j}, & x_{2j}, & \ldots, & x_{l-1j}, & -, & -, & x_{l+2j,} & \ldots & x_{nj} \\ 1, & 2, & \ldots, & l-1, & l, & l+1, & l+2, & \ldots & n. \end{array}$$

Понятно, что пропущенные значения можно заменить средними или расчетными, предварительно построив для этих целей трендовую модель. Рассмотрим более интересный подход, основанный, как и предыдущий метод, на идее адаптивных ожиданий, но в отличие от него в этом подходе будет использоваться модифицированный вариант формулы адаптивных ожиданий.

Запишем фрагмент временного ряда

$$x_{l-2\,j},\ x_{l-1\,j},\ \tilde{x}_{lj}$$

который имел бы место после восстановления. Для данных этого фрагмента можно записать

$$x_{l-1\,j} = \alpha\, x_{l-2\,j} + (1-\alpha)\,\tilde{x}_{lj},\ 0 < \alpha < 1$$

или

$$\tilde{x}_{lj} = \frac{1}{1-\alpha} x_{l-1\,j} - \frac{\alpha}{1-\alpha} x_{l-2\,j},\ 0 < \alpha < 1.$$

Аналогично для второго пропуска

$$x_{l+2\,j} = \alpha\,\tilde{x}_{l+1j} + (1-\alpha)\, x_{l+3\,j},\ 0 < \alpha < 1$$

или

$$\tilde{x}_{l+1\,j} = \frac{1}{\alpha} x_{l+2\,j} - \frac{1-\alpha}{\alpha} x_{l+3\,j},\ 0 < \alpha < 1.$$

На следующем шаге по данным с ожидаемыми значениями пропусков

$$\begin{array}{c} x_{1j},\ x_{2j},\ \ldots,\ x_{l-1j},\ \tilde{x}_{lj},\ \tilde{x}_{l+1j}\ ,\ x_{l+2j},\ \ldots\ x_{nj} \\ 1,\ \ 2,\ \ \ldots,\ \ l\text{-}1,\ \ \ l,\ \ \ l\text{+}1,\ \ l\text{+}2,\ \ \ldots,\ \ n \end{array}$$

строится трендовая модель $f(t,\ \alpha)$, которая после настройки параметра α используется для определения восстанавливаемых значений, т.е.

$$\tilde{x}_{lj} = f(l,\ \alpha^*) \quad \text{и} \quad \tilde{x}_{l+1\,j} = f(l+1,\ \alpha^*).$$

Развивая идею применения адаптивных ожиданий для восстановления пропусков, рассмотрим ситуацию, когда в данных имеется три рядом расположенных пропуска на $l-1$, l, $l+1$ местах. Тогда формулы для адаптивных ожиданий, которыми заменяются пропуски, представим в виде

$$\tilde{x}_{l-1\,j} = \frac{1}{1-\alpha} x_{l-2\,j} - \frac{1}{1-\alpha} x_{l-3\,j},\ 0 < \alpha < 1,$$

$$\tilde{x}_{l+1\,j} = \frac{1}{\alpha} x_{l+2\,j} - \frac{1-\alpha}{\alpha} x_{l+3\,j},\ 0 < \alpha < 1,$$

$$\tilde{x}_{lj} = \alpha\,\tilde{x}_{l-1\,j} + (1-\alpha)\ \tilde{x}_{l+1\,j},\ 0 < \alpha < 1.$$

Таким образом, нами рассмотрены практически все ситуации, которые могут встретиться в задачах анализа и использования данных при построении различного рода математических моделей.

Изложенные здесь методы главным образом применимы для восстановления пропусков в независимых (факторных) переменных. В принципе эти методы можно использовать и для восстановления пропусков в зависимой переменной, но есть другие, более эффективные методы, ориентированные на решение именно таких задач.

Рассмотренные методы являются составной частью современного аппарата экономического прогнозирования. Их применение повышает надежность прогнозных расчетов в тех ситуациях, когда в данных обнаруживаются неточности и искажения, связанные с дефектами формирования выборочных совокупностей.

ЛИТЕРАТУРА

1. **Бочаров В.В.** Финансовое моделирование. – СПб.: Питер, 2000.

2. **Булгакова И.Н.** Адаптивно-имитационное моделирование прогнозных оценок предкризисных ситуаций // Энергия. – 2001. – № 4 (41). – С. 100–105.

3. **Городничев П.Н., Городничева К.П.** Финансовое и инвестиционное прогнозирование: Учебное пособие. – М.: Экзамен, 2005.

4. **Магнус Я.Р.** Эконометрика. Начальный курс: Учебник. – М.: Дело, 2004.