УДК 65.011.8

А.Д. Цветков
(научный руководитель – проф. ВШБИ С.Г. Светуньков)
Санкт-Петербургский политехнический университет Петра Великого

РАЗРАБОТКА АЛГОРИТМА ЗАДАНИЯ Z-МНОЖИТЕЛЕЙ ДЛЯ АДАПТАЦИИ МОДЕЛИ К СВОЙСТВАМ МОДЕЛИРУЕМОГО ПРОЦЕССА

Актуальность. Современные экономисты часто сталкиваются с задачей построения моделей прогнозирования социально-экономической динамики. Оценки коэффициентов таких моделей можно построить различными способами: от МНК до метода спейсингов. Одним из альтернативных методов является метод z-множителей. Правильный подбор вида функций z-множителей обеспечивает наиболее точную оценку коэффициентов модели, описывающей необратимые процессы.

Целью работы является предложение алгоритма, позволяющего подобрать z-множители для оценки линейных прогнозных моделей с учетом имеющихся данных – выборки наблюдений.

Задачи исследования:

– рассмотреть математическую постановку задачи о подборе z-множителей для линейных моделей;

– проанализировать возможность применения к поставленной задаче методов символьной регрессии;

– разработать алгоритм подбора z-множителей на примере оценок коэффициентов линейной модели прогнозирования.

Результаты.

1. Пусть $t \in T$ – множество индексов элементов обучающей выборки. Пусть также известны одномерные данные: $x_t \in X, y_t \in Y; \varepsilon_t \in \mathrm{E};\ X, Y, \mathrm{E} \subset \mathbb{R}$.

Рассмотрим прогнозную модель следующего вида:

$$\hat{y}_t = a_1 x_t + a_0 + \varepsilon_t \tag{1}$$

Модель (1) является линейной однофакторной прогнозной моделью. Для такой модели известен способ оценивания ее коэффициентов a_1, a_0 с помощью метода z-множителей [1]:

$$\begin{cases} \sum_{t\in T} y_t z_{0t}(x) = a_0 \sum_{t\in T} z_{0t}(x) + a_1 \sum_{t\in T} x_t z_{0t}(x) + \sum_{t\in T} \varepsilon_t z_{0t} \\ \sum_{t\in T} y_t z_{1t}(x) = a_0 \sum_{t\in T} z_{1t}(x) + a_1 \sum_{t\in T} x_t z_{1t}(x) + \sum_{t\in T} \varepsilon_t z_{1t} \end{cases} \tag{2}$$

В системе уравнений (2) примем ошибки аппроксимации ε_t равными нулю, поскольку реальные данные процессов социально-экономической динамики не предполагают наличие сведений об ошибках. Теперь в данной системе неизвестными являются виды функций z_{0t}, z_{1t}, а также оцениваемые коэффициенты a_1, a_0.

Пусть также $i \in I$ – множество индексов элементов тестовой выборки.

Тогда задачу прогнозирования можно свести к минимизации среднеквадратичной ошибки аппроксимации:

$$\sum_{i\in I} (y_i - a_1 x_i - a_0)^2 \to min \tag{3}$$

Выразим в явном виде уравнения отыскиваемых коэффициентов модели из системы (2):

$$a_1 = \frac{(\sum_t y_t z_{0t}(x) \sum_t z_{1t}(x) - \sum_t y_t z_{1t}(x) \sum_t z_{0t}(x))}{(\sum_t x_t z_{0t}(x) \sum_t z_{1t}(x) - \sum_t x_t z_{1t}(x) \sum_t z_{0t}(x))} \tag{4}$$

$$a_0 = \frac{(\sum_t y_t z_{1t}(x) \sum_t x_t z_{0t}(x) - \sum_t y_t z_{0t}(x) \sum_t x_t z_{1t}(x))}{(\sum_t z_{1t}(x) \sum_t x_t z_{0t}(x) - \sum_t z_{0t}(x) \sum_t x_t z_{1t}(x))} \quad (5)$$

С учетом выражений (4), (5) можно переформулировать задачу (3).

Необходимо найти $z_{0t}(x), z_{1t}(x)$, такие, что

$$\sum_{i \in I} \left(y_i - \frac{\sum_t y_t z_{0t}(x) \sum_t z_{1t}(x) - \sum_t y_t z_{1t}(x) \sum_t z_{0t}(x)}{\sum_t x_t z_{0t}(x) \sum_t z_{1t}(x) - \sum_t x_t z_{1t}(x) \sum_t z_{0t}(x)} x_i + \frac{\sum_t y_t z_{1t}(x) \sum_t x_t z_{0t}(x) - \sum_t y_t z_{0t}(x) \sum_t x_t z_{1t}(x)}{\sum_t z_{1t}(x) \sum_t x_t z_{0t}(x) - \sum_t z_{0t}(x) \sum_t x_t z_{1t}(x)} \right)^2 \to min \quad (6)$$

2. Обратимся теперь к задаче символьной регрессии: она заключается в нахождении математического выражения в символьной форме, отражающего зависимость между представленным набором независимых факторов (x_t) и соответствующими им значениями зависимых переменных (y_t). [2]

Пусть имеется выборка $x_i \in X \subset \mathbb{R}^m,\ y_i \in Y \subset \mathbb{R}: (x_i,\ y_i),\ i = \overline{1,\ n}$, $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$.

$$y_i = f(x_i) \quad (7)$$

Необходимо найти выражение $\hat{f}(x_i)$ в символьном виде, которое аппроксимирует зависимость f между x_i и y_i. При этом задается функционал качества:

$$S(x, y, f) \to min \quad (8)$$

3. Неожиданно заметим, что выражение (6) в задаче подбора z-множителей представляет собой функционал качества из выражения (8), а искомая оценка $\hat{f}$ для выражения (7) представима в виде

$$\hat{y}_t = \frac{\sum_t y_t z_{0t}(x) \sum_t z_{1t}(x) - \sum_t y_t z_{1t}(x) \sum_t z_{0t}(x)}{\sum_t x_t z_{0t}(x) \sum_t z_{1t}(x) - \sum_t x_t z_{1t}(x) \sum_t z_{0t}(x)} x_i + \frac{\sum_t y_t z_{1t}(x) \sum_t x_t z_{0t}(x) - \sum_t y_t z_{0t}(x) \sum_t x_t z_{1t}(x)}{\sum_t z_{1t}(x) \sum_t x_t z_{0t}(x) - \sum_t z_{0t}(x) \sum_t x_t z_{1t}(x)} \quad (9)$$

Заметим также, что вид аппроксимации зависимости $\hat{f}$ в исходной задаче нам практически известен, что подтверждает уравнение (9), за исключением вида двух z-множителей. Чтобы привести задачу к такой, которая бы решилась методами символьной регрессии, вернемся к системе (2) и сложим два уравнения системы (с учетом обнуления значений ошибок аппроксимации):

$$\sum_{t \in T} y_t \big(z_{0t}(x) + z_{1t}(x)\big) = a_0 \sum_{t \in T} \big(z_{0t}(x) + z_{1t}(x)\big) + a_1 \sum_{t \in T} x_t \big(z_{0t}(x) + z_{1t}(x)\big) \quad (10)$$

В выражении (10) оба z-множителя вносят одинаковый вклад в нахождение коэффициентов модели и представляют собой суперпозицию: $Z(x) = z_{0t}(x) + z_{1t}(x)$. Поэтому зафиксируем вид одного из множителей заранее и будем искать символьное выражение только второго множителя.

Тогда необходимо найти $z(x)$, такой, что:

$$\sum_{i \in I} \left(y_i - \frac{\sum_t y_t \varphi(x) \sum_t z(x) - \sum_t y_t z(x) \sum_t \varphi(x)}{\sum_t x_t \varphi(x) \sum_t z(x) - \sum_t x_t z(x) \sum_t \varphi(x)} x_i + \frac{\sum_t y_t z(x) \sum_t x_t \varphi(x) - \sum_t y_t \varphi(x) \sum_t x_t z(x)}{\sum_t z(x) \sum_t x_t \varphi(x) - \sum_t \varphi(x) \sum_t x_t z(x)} \right)^2 \to min \quad (11)$$

где вид $\varphi(x)$ – фиксирован.

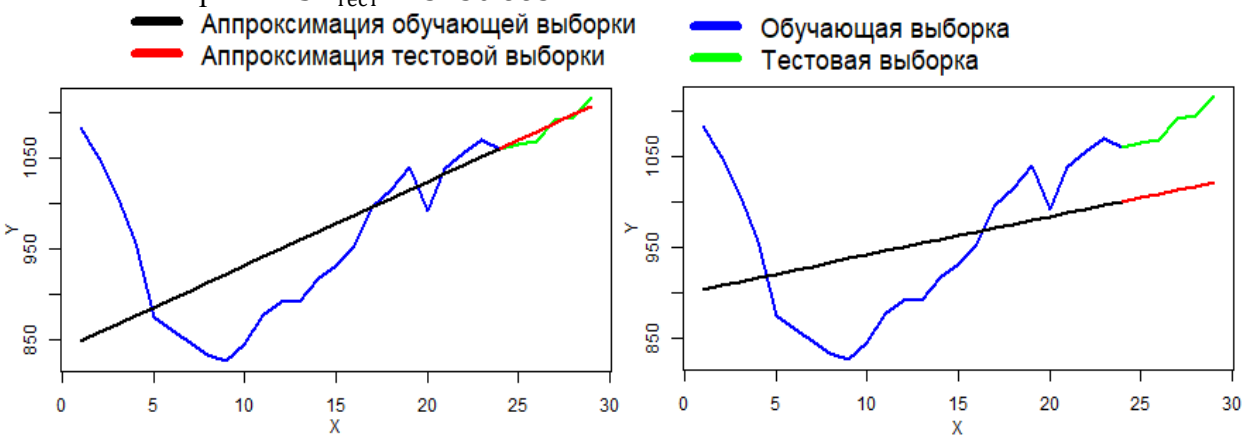
4. При решении поставленной задачи (11) символьной регрессии будем пользоваться методом грамматической эволюции (генетическим алгоритмом). Функция представляется в виде синтаксического дерева, в листьях которого находятся независимые переменные и числа, а в узлах – унарные и бинарные операции [3].

Метод грамматической эволюции заключается в следующем: на первой итерации генерируется лес (популяция) случайных символьных деревьев. Затем из популяции выбираются наилучшие – дающие наименьшее значение функционала качества (11). На их основе с помощью генетических операций, таких как скрещивание (обмен случайно выбранных поддеревьев), мутации (замена функции в узле на другую, случайно выбранную; замена поддерева на случайно сгенерированное; удаление промежуточного унарного узла; создание нового корня дерева; перестановка дочерних поддеревьев для некоммутативных операций) генерируется новая популяция деревьев. Из них снова выбираются наилучшие по функционалу качества. Процесс продолжается до тех пор, пока возможно уменьшение значения минимизируемой функции. Процесс является сходящимся, поскольку на каждой итерации выбирается наилучшая популяция деревьев.

5. На языке программирования R [4] была разработана программа, реализующая генетический алгоритм: представление функции в виде синтаксического дерева (глубина дерева ≤5); мутации и скрещивание; итеративный выбор популяций. В качестве исходных данных для тестирования разработанного алгоритма были выбраны сведения о производстве электроэнергии, млрд кВт·ч, 1990-2018г. Выборка была поделена на тестовую и обучающую с коэффициентом $learnratio = 0.8$.

Результаты работы программы графически изображены на рис. 1а. Первый z-множитель был задан как $z_{0t}(x) = 1$. Найденный программно $z_{1t} = x^2 \cos x$. Полученное уравнение аппроксимации имеет вид $\hat{y}_t = 840.2480 + 9.1641x_t$. Среднеквадратичная ошибка на всей выборке $MSE_{\text{общ}} = 5687.248$, на тестовой выборке - $MSE_{\text{тест}} = 41.283$.

Дополнительно было проведено исследование исходных данных с помощью МНК-оценок: известно, что при выборе $z_{0t} = 1, z_{1t} = x_t$ система (2) будет соответствовать системе нормальных уравнений МНК [5]. Результаты использования МНК-оценок приведены на рис. 1б. Полученное уравнение аппроксимации имеет вид $\hat{y}_t = 900.0937 + 4.177x_t$. Среднеквадратичная ошибка по всей выборке составила $MSE_{\text{общ}} = 5892.983$, на тестовой выборке - $MSE_{\text{тест}} = 5230.665$.



а) б)

Рис. 1. Линейная аппроксимация исходных данных: а) с помощью оценок разработанного алгоритма поиска z-множителей; б) с помощью МНК-оценок

Выводы. Разработанный алгоритм подбора z-множителей на основе грамматической эволюции оказался применимым к решению реальных задач: использование грамматической эволюции позволило снизить ошибку на тестовых данных в 127 раз по сравнению с МНК-оценками. Необходимость выбора z-множителей с учетом особенностей исходных данных была подтверждена.

ЛИТЕРАТУРА

1. Светуньков С. Г. Эконометрические методы прогнозирования спроса (на примере промышленной энергетики). М.: Изд-во МГУ, 1993. С. 86.
2. Хритоненко Д. И. Адаптивные коллективные нейро-эволюционные алгоритмы интеллектуального анализа данных. Диссертация. Красноярск, 2017. [Рукопись]. С. 15.
3. Символьная регрессия [Электронный ресурс]. URL: https://habr.com/ru/post/163195/ (дата обращения: 18.11.2021)
4. The R Project for Statical Computing [Электронный ресурс]. URL: https://www.r-project.org/ (дата обращения: 18.11.2021)
5. Светуньков С.Г., Светуньков И.С. Методы социально-экономического прогнозирования: Учебник для вузов. Том II. – СПб.: Изд-во СПбГУЭФ, 2010. С. 63.