УДК 519.237.5

Методические и инструментальные средства построения некоторых типов регрессионных моделей

М.П. Базилевский[1*], С.И. Носков[1]

[1]Иркутский государственный университет путей сообщения. Чернышевского15, Иркутск, Россия.
Статья поступила 14.12 2011, принята 15.02.2012

Статья посвящена проблеме выбора формы связи между независимыми переменными в регрессионной модели. Кратко рассмотрены недостатки существующих методов и основного программного обеспечения для выбора таких форм. Предложена и подробно рассмотрена технология организации «конкурса» моделей, которая заключается в формировании множества альтернативных вариантов регрессий и последующем выборе лучшей из них с использованием многокритериального подхода. В рамках «конкурса» предложены и рассмотрены четыре формы регрессионных зависимостей: аддитивная, с использованием эффекта запаздывания, с преобразованием зависимой переменной и линейно-мультипликативная. Рассмотренная технология реализована в новой версии программного комплекса автоматизации процесса построения регрессионных моделей. К основным возможностям данного программного продукта относятся: удобство работы с исходными данными, ручной режим работы, четыре автоматических режима работы, интерпретация результатов моделирования, прогнозирование по модели. К достоинствам комплекса относится также то, что он позволяет строить только те модели, которые соответствуют «физическому» смыслу независимых факторов. Для демонстрации работы приводится пример решения задачи с классическими данными по работе выпарного аппарата на большом промышленном предприятии с использованием данного программного комплекса. Полученная таким образом модель оказалась лучше классической. Реализованные в новой версии программного комплекса автоматизации процесса построения регрессионных моделей методы и алгоритмы не имеют аналогов в других программных продуктах, а его применение при проведении регрессионного анализа является весьма эффективным. Эта система может быть использована практически для любой предметной области, в которой имеются количественные статистические данные.

Ключевые слова: регрессионный анализ, «конкурс» моделей, критерии адекватности, программный комплекс.

Methodology and instrumental tools for construction some types of regression models

M.P. Bazilevskiy[1*], S.I. Noskov[1]

[1]Irkutsk State University of Railway Engineering, 15,Chernyshevskogo str., Irkutsk, Russia
Received 14.12.2011; Accepted 15.02.2012

This article is devoted to the problem of the communication form choice between independent variables in the regression model. The limitations of existing methods and basic software for the selection of such forms are briefly discussed. In detail, the organization of models «competition» technology has been proposed and discussed. It is based on producing a set of alternative regressions and the subsequent selection the best one using a multi-criteria approach. Under the «competition», four forms of regression have been proposed and discussed: additive, using the delay effect, the transformation of the dependent variable, and linear-multiplicative. The considered technology has been implemented in the new version of the software package for automation construction process of regression models. The main features of this software are: usability of the initial data, manual operation mode, four automation operation modes, interpretation of modeling results, forecasting by the model. One of the advantages of the complex is the fact that it allows us to construct only those models that meet the «physical» sense of independent factors. To demonstrate its performance, an example of solving the problem with classical data on the evaporator operation at a large industrial plant using this software package is proposed. Thus, the obtained model proved to be better than the classical one. Being implemented in the new version of software package for automation construction process of regression models, the methods and algorithms don't have any analogs in other software products, its application in the course of regression analysis being very effective. This system can be used practically in any subject area which deals with statistical data.

Keywords: regression analysis, models «competition», adequacy tests, software package.

В настоящее время в связи с бурным развитием информационных технологий в различных областях знаний накоплено большое количество информации в виде числовых данных. Одним из наиболее популярных инструментов анализа этих данных является регрессионный анализ [1, 2]. Традиционно при проведении регрессионного анализа предполагается наличие двух различных групп анализируемых факторов – зависимых и независимых переменных. Сущность регрессионного анализа заключается в выявлении степени влияния

* E-mail address: mik2178@yandex.ru

второй группы на первую и оценке значений неизвестных параметров модельных конструкций.

Одной из основных проблем, связанных с построением регрессионных моделей, является выбор наилучшей формы связи между переменными. Обзор различных таких форм можно найти в многочисленной литературе (см., например, [3-8]). Проблема выбора формы модели также тесно связана с проблемой выбора наилучшего набора независимых переменных, известной в зарубежной литературе, как «subset selection in regression» [9]. К сожалению, к настоящему времени не изобретено универсальных аналитических или итерационных методов, гарантирующих построение оптимальных по форме регрессий. Но стоит отметить, что все же существует немалое количество эвристических процедур, реализация которых приводит к построению «хороших» моделей (см., например, [2-4], [9-11]). К ним относятся пошаговые процедуры (stepwise) включения, исключения, включения-исключения, алгоритмы последовательной замены и др. Такие процедуры не гарантируют построение оптимальной модели в рамках конкретной задачи, а иногда даже приводят к проблемным ситуациям [10, 12]. Как отдельный класс можно выделить методы индуктивного порождения регрессионных моделей [13, 14]. Такие методы позволяют быстро построить «хорошую» модель, но она зачастую получается переобученной и совершенно не интерпретируемой по смыслу. Проведенный обзор литературы позволяет сделать вывод, что на сегодняшний день для построения оптимальных в заданном классе уравнений регрессий и по заданным критериям адекватности необходимо применение процедур переборного характера. Но также этот способ и самый длительный из всех и поэтому требует привлечения в процесс моделирования вычислительных способностей современных компьютеров.

В настоящее время существует очень большой выбор программного обеспечения для статистических исследований (см., например, [15-21]). Согласно [15], основную часть имеющихся статистических пакетов составляют специализированные пакеты и пакеты общего назначения. Специализированные пакеты содержат методы из одного или двух разделов статистики или методы, используемые в конкретной предметной области (например, пакеты для регрессионного анализа). Пакеты общего назначения характеризуются отсутствием прямой ориентации на конкретную предметную область и имеют широкий диапазон статистических методов.

Рассмотрим достоинства и недостатки основных статистических пакетов с позиций регрессионного анализа, а также наличие в них функций для решения задачи выбора формы связи между независимыми переменными. Самыми популярными пакетами общего назначения являются SPSS, STATISTICA, SAS, STATGRAPHICS, STADIA. К ним также можно отнести множество эконометрических систем, таких, как AREMOS, AUTOBOX, B34S, ВЕТАНАТ, CEF, EasyReg International, EVIEWS, FP, GRETL, IDIOM, LIMDEP, MICROFIT, Modeleasy+, MODLER, MODLER BLUE, MOSAIC, NLOGIT, PcGets, PcGive, PcNaive, RATS, REG-X, SHAZAM, SORITEC, STAMP, STATA, TROLL, TSP, WinSolve, WYSEA. Следует отметить, что практически в каждом из перечисленных пакетов имеется целый арсенал инструментов для проведения статистических исследований, подробное руководство пользователя, а также мощный графический интерфейс. К недостаткам же следует отнести довольно стандартный набор функций для проведения регрессионного анализа: формирование матрицы коэффициентов корреляции, построение линейной множественной регрессии пошаговым методом, вычисление критерия множественной детерминации модели и др. Хотя в некоторых пакетах и содержатся встроенные языки программирования, но для их освоения требуется немало времени.

Из специализированных пакетов для регрессионного анализа были рассмотрены SYSTAT TableCurve 2D v5.01, SYSTAT TableCurve 3D v4.0, DataFit 9.0, LabFit v.7.2.43, MVR Composer (см. обзор в [21]). Они содержат уже гораздо большее количество функций для построения регрессий, таких, как автоматический выбор лучшей регрессии из заданного множества альтернатив, использование различных методов оценивания неизвестных параметров и т. д. Но среди специализированных пакетов для регрессионного анализа можно выделить и ряд недостатков.

Отсутствие возможности работать с многомерными данными. Например, пакеты SYSTAT TableCurve содержат встроенные библиотеки из тысяч только двухмерных и трехмерных моделей, а система DataFit хоть и позволяет работать с шестью независимыми переменными, но ее библиотека уравнений настолько малочисленна, что добавлять структуры регрессий приходится вручную.

Отсутствие многокритериального выбора. Все пакеты осуществляют поиск только по одному критерию и затем упорядочивают модели по степени адекватности.

Построение моделей, которые не соответствуют смыслу факторов. Очень часто строятся полиномы, которые достаточно точно аппроксимируют данные, но такие модели не могут быть интерпретированы.

Ранее для решения задачи построения регрессионных моделей авторами был разработан соответствующий программный комплекс, подробное описание которого приведено в [22]. С учетом проведенного анализа, а также с разработкой новых алгоритмов и функций построения моделей возникла необходимость развития и реинжиниринга данного комплекса. В данной работе представлено описание возможностей и пример работы с новой версией программного комплекса автоматизации процесса построения регрессионных моделей (ПК АППРМ). Основным ядром этого комплекса является технология организации «конкурса» моделей [23], которая состоит в построении множества альтернативных вариантов регрессий и последующем выборе лучшей из них, исходя из наличия векторного критерия оценки адекватности каждого из вариантов. Рассмотрим подробнее эту технологию.

Пусть дано линейное многофакторное регрессионное уравнение:

$$y_k = \alpha_0 + \sum_{i=1}^{m} \alpha_i x_{ki} + \varepsilon_k, \quad k = \overline{1,n}, \qquad (1)$$

где n – число наблюдений (длина выборки); y_k и x_{ki}, $k=\overline{1,n}$, $i=\overline{1,m}$ – значения зависимой и независимых переменных соответственно; α_i, $i=\overline{1,m}$ – подлежащие оцениванию параметры; ε_k, $k=\overline{1,n}$ – ошибки аппроксимации.

Присутствие в уравнении (1) ошибок аппроксимации означает, что данная связь описывает процесс не точно, а с некоторой погрешностью. Это может быть вызвано:

- неточностями в регистрации значений зависимой и независимых переменных;
- влиянием помех;
- неучетом ряда значимых факторов;
- неточным (неудачным) выбором формы связи

между переменными или метода оценивания параметров.

Уравнение (1) может быть использовано для тех или иных целей после вычисления вектора параметров α. Для нахождения неизвестных параметров можно использовать, например, методы наименьших квадратов (МНК), модулей (МНМ) и другие.

Определимся теперь с одним важным методологическим обстоятельством, связанным с трактовкой ε. Существует два подхода к интерпретации исходных данных и их статистической обработке [23]. В соответствии с первым из них совокупность рядов наблюдений трактуется как выборка из соответствующей так называемой генеральной совокупности, как ее представитель, по свойствам которого можно судить о свойствах всей совокупности. При такой трактовке исследователь основывает свои методы исследования последних на вероятностной природе исходных данных, используя для этого соответствующую вероятностную модель. В этом случае ε в (1) можно считать распределенной по некоторому закону случайной величиной и основывать на этом анализ ее свойств.

При втором подходе исследователь имеет дело со своего рода одной, уникальной выборкой (обычно весьма ограниченной), отражающей наблюдения за некоторым реальным объектом (что и имеет часто место при моделировании социально-экономических и других систем), а всякие априорные сведения о вероятностной природе исходных данных отсутствуют. Все полученные в данной работе результаты находятся в рамках второго подхода, при котором ε в (1) трактуется как ошибка аппроксимации, и только.

Большинство реальных социально-экономических процессов не получается адекватно описывать линейной связью (1), поэтому возникает необходимость в построении более сложных моделей. Построим посредством варьирования вида аппроксимирующей функции и набора независимых переменных с учетом их преобразований и комбинаций множество из r вариантов $\mathcal{M}=\{M_1,M_2,\ldots,M_r\}$, среди которых нужно выбрать наиболее приемлемый, руководствуясь значениями критериев $K_1,K_2,\ldots,K_l$ для каждого из вариантов, то есть матрицей $\mathcal{K}=\|K_i(M_j)\|$, $i=\overline{1,l}$, $j=\overline{1,r}$. В данной работе использовались следующие критерии адекватности:

Критерий множественной детерминации R, выражающий степень согласованности вычисленных и фактических значений зависимой переменной:

$$R=\frac{\sum_{k=1}^{n}(\tilde{y}_k-\bar{y})^2}{\sum_{k=1}^{n}(y_k-\bar{y})^2},$$

где y_k – фактические значения зависимой переменной, $\tilde{y}_k$ – вычисленные значения, $\bar{y}=\frac{1}{n}\sum_{k=1}^{n}y_k$ – среднее значение.

Критерий Фишера F, который указывает на значимость критерия детерминации:

$$F=\frac{R}{1-R}\cdot\frac{n-m}{m}.$$

Величина остаточной дисперсии S, определяющая меру вариации выходного показателя относительно линии регрессии:

$$S=\frac{1}{n-m}\sum_{k=1}^{n}\varepsilon_k^2.$$

Средняя относительная ошибка аппроксимации E, указывающая на точность модели и обычно применяемая в инженерных расчетах:

$$E=\frac{1}{n}\sum_{k=1}^{n}\left|\frac{y_k-\tilde{y}_k}{y_k}\right|\cdot 100\%.$$

Критерий Дарбина-Уотсона DW, указывающий на наличие или отсутствие автокорреляции (положительной или отрицательной) остатков ε_k:

$$DW=\frac{\sum_{k=2}^{n}(\varepsilon_k-\varepsilon_{k-1})^2}{\sum_{k=1}^{n}\varepsilon_k^2}.$$

Идеальное значение критерия DW, указывающее на полное отсутствие автокорреляции остатков, равно двум.

Для удобства приведем все критерии к однородному виду. Будем считать, что для всех $i=\overline{1,l}$ лучшим вариантом по i-му критерию является тот, который соответствует максимальному элементу i-ой строки матрицы $\mathcal{K}$. Для этого элементам столбцов, соответствующих критериям оценки остаточной дисперсии и средней относительной ошибки аппроксимации, следует приписать знак «минус», так как известно, что

$$\min_{M\in\mathcal{M}} K_i(M) = -\max_{M\in\mathcal{M}}(-K_i(M)).$$

Поскольку критерий DW принимает значение в интервале [0;4], и лучшим его значением является 2, следует преобразовать DW к виду

$$DW^* = \begin{cases} DW, & при\ DW(M) \le 2 \\ 4 - DW, & при\ DW(M) > 2 \end{cases}.$$

В теории принятия решений разработано большое количество эффективных алгоритмов решения многокритериальных задач, многие из которых вполне применимы и при выборе лучшего варианта регрессионной зависимости. Так как в нашем случае ЛПР (лицо, принимающее решение) может не владеть никакой информацией о сравнительной значимости критериев адекватности, в этом случае рационально использовать метод «идеальной» точки [24], идея которого состоит в следующем.

Прежде всего, элементы матрицы $\mathcal{K}$ нормируются по правилу:

$$\tilde{K}_i(M_j) = \frac{K_i(M_j) - K_i^-}{K_i^+ - K_i^-},\ i = \overline{1,l},\ j = \overline{1,r},$$

где $K_i^- = \min_{M\in\mathcal{M}} K_i(M)$, $K_i^+ = \max_{M\in\mathcal{M}} K_i(M)$.

Затем определяются максимальные элементы K_i^*, $i = \overline{1,l}$ в каждой строке матрицы $\tilde{K}$:

$$K_i^* = \max_{j=1,r} \tilde{K}_i(M_j).$$

Таким образом, «идеальная» точка $K^* = (K_1^*, K_2^*, \ldots, K_l^*)$ представляет собой вектор, каждая компонента которого равна максимальному значению соответствующего критерия. Для реальных задач многокритериального выбора лучшего варианта регрессионного уравнения обычно отсутствует альтернатива, доставляющая максимум всем критериям одновременно. Поэтому метод «идеальной» точки предполагает поиск альтернативы, образ которой в критериальном пространстве наиболее близок в некоторой метрике (например, евклидовой) к точке K^*:

$$M^* = \arg\min_{M\in\mathcal{M}} \sum_{i=1}^{l} \left(K_i^* - \tilde{K}_i(M)\right)^2.$$

Рассмотрим теперь способы формирования множества $\mathcal{M}$. В данной работе авторами предложены следующие формы регрессионных зависимостей.

Аддитивная по параметрам регрессия.

Пусть для модели (1) задано p регрессоров и множество независимых переменных $X = \{x_1, x_2, \ldots, x_m\}$. Введем набор преобразований для каждой такой переменной $F(x) = \{f_1(x), f_2(x) \ldots f_l(x)\}$. В качестве f_j используются следующие элементарные функции:

а) *показательная* функция $f_j = \beta^{\alpha\cdot x}$, $\beta > 1$;

б) *степенная* функция $f_j = x^{\beta}$;

в) *логарифмическая* функция $f_j = \log_{\beta} x$, $\beta > 1$;

г) *тригонометрические* функции $f_j = \sin(\beta \cdot x)$, $f_j = \cos(\beta \cdot x)$.

Смысл введения таких преобразований состоит в расширении исходного набора объясняющих переменных с целью последующего выбора в новом, расширенном наборе из $m \cdot l$ переменных совокупности из p наиболее «информативных» факторов. Тем самым осуществляется переход от линейной регрессии (1) со свободным членом к нелинейной по факторам, но линейной по параметрам аддитивной зависимости вида:

$$y_k = \alpha_0 + \sum_{h=1}^{p} \alpha_h f_{ji}(x_{ki}) + \varepsilon_k, \qquad (2)$$

где f_{ji} – преобразование с номером j для i-ой объясняющей переменной. Общее число возможных моделей для данного алгоритма равно

$$r = C_{m\cdot l}^{p}.$$

Регрессия с запаздыванием.

По аналогии с аддитивной регрессией, для формирования множества альтернативных вариантов моделей можно использовать лаги запаздывания для независимых переменных. Введем в рассмотрение набор таких лагов $\{\tau_0, \tau_1, \ldots, \tau_l\}$. Тогда в общем виде полученные регрессии можно представить так:

$$y_t^k = \alpha^{(0)} + \sum_{h=1}^{p} \alpha^{(h)} x_{t-\tau_j}^{(ki)} + \varepsilon_t^{(k)}, \qquad (3)$$

где x_t^k, y_t^k – значения переменных в момент времени t, τ_j – лаг запаздывания с номером j. Общее число моделей для данного алгоритма равно $r = C_{m\cdot(l+1)}^{p}$.

Линейно-мультипликативная регрессия (ЛМР).

Данный алгоритм подробно изложен в [25]. Для формирования множества альтернативных вариантов моделей используются комбинации произведений независимых переменных. Тогда в общем случае все возможные линейно-мультипликативные регрессии можно представить в виде:

$$y_k = \alpha_0 + \sum_{i=1}^{p} \alpha_i \prod_{j=1}^{m} x_{kj}^{\sigma_{sji}} + \varepsilon_k, \qquad (4)$$

$$s = \overline{1,r},\ k = \overline{1,n}$$

где p – число регрессоров; σ_{sji} – булева переменная, задаваемая по правилу:

$$\sigma_{sji} = \begin{cases} 1, & \text{если в } s-\text{ой регрессии } j-\text{ая} \\ & \text{переменная } x_j \text{ входит в } i-\text{ое слагаемое} \\ 0, & \text{в противном случае} \end{cases}$$

Авторами предложено три стратегии задания характера вхождения независимых переменных в модель (4):

а) не требуется обязательного вхождения каждой независимой переменной в ЛМР. Общее число моделей для данной стратегии

$$r = C^{p}_{2^m-1};$$

б) каждая независимая переменная входит в ЛМР только 1 раз. Общее число моделей

$$r = \frac{1}{p!}\sum_{i=0}^{p}(-1)^i C_p^i (p-i)^m;$$

в) каждая независимая переменная входит в ЛМР хотя бы 1 раз. Общее число моделей

$$r = \sum_{i=0}^{m-1}(-1)^i \cdot C_m^i \cdot C^{p}_{2^{m-i}-1}.$$

Регрессия с преобразованием зависимой переменной.

Данный способ также во многом схож с алгоритмом формирования аддитивных регрессий. В нем использованы преобразования зависимой переменной в виде элементарных функций, но такие модели содержат уже не аддитивную, а мультипликативную ошибку ε_k:

$$y_k = f_j(\alpha_0 + \sum_{h=1}^{p}\alpha_h x_{ki} + \varepsilon_k), \qquad (5)$$

где f_j – преобразование с номером j.

Это связано с тем, что оценивание неизвестных параметров α в случае с аддитивной ошибкой приведет к использованию времяемких итеративных процедур – методам Левенберга-Марквардта, Гаусса-Ньютона и т. д. А в случае с мультипликативной ошибкой модели типа (5) можно свести к линейным по факторам и оценивать неизвестные параметры α хорошо известными методами наименьших квадратов или наименьших модулей.

Новая версия ПК АППРМ реализована на языке программирования С++. К основным возможностям комплекса относятся следующие.

Удобство работы с исходными данными.

Исходные данные можно легко ввести в систему вручную, либо импортировать их, например, из текстового файла. Предусмотрена возможность изменения этих данных или полное их удаление.

Ручной режим работы.

Этот режим позволяет пользователю выбрать практически любую форму связи между независимыми переменными в модели. Структура модели представлена в системе в виде иерархического дерева, управляя узлами которого, можно задать практически любую зависимость.

Автоматические режимы работы.

Эти режимы основаны на технологии «конкурса» моделей. В системе реализовано четыре автоматических алгоритма, соответствующих рассмотренным ранее классам регрессий: аддитивная, мультипликативная, с запаздыванием, с преобразованием отклика. Пользователю необходимо задать лишь начальные параметры: алгоритм, зависимую и независимые переменные, метод оценивания, число регрессоров, критерии адекватности и др., после чего система сформирует множество альтернативных вариантов регрессий и выберет из них лучшую. В ПК АППРМ возможно строить уравнения, содержащие не более 6 регрессоров, и в зависимости от заданных параметров, комплекс может строить десятки миллионов альтернативных регрессий. Выбор лучшего уравнения можно осуществлять одновременно по пяти критериям адекватности. С целью получения интерпретируемых моделей (соответствующих смыслу факторов) в систему встроена интеллектуальная подсистема отсева бессмысленных уравнений.

1. *Интерпретация результатов моделировния.*

При выборе любой построенной модели в системе автоматически формируется подробный анализ результатов моделирования. Он содержит уравнение модели, критерии адекватности, а также, благодаря встроенной подсистеме интерпретации моделей, некоторые советы и рекомендации по применению этой зависимости на практике.

2. *Прогнозирование по модели.*

Предусмотрены два типа прогнозов: точечные и интервальные. После проведения процесса прогнозирования полученные прогнозные значения отображаются на графике значений зависимой переменной.

Интерфейс программного комплекса представлен на рис. 1.

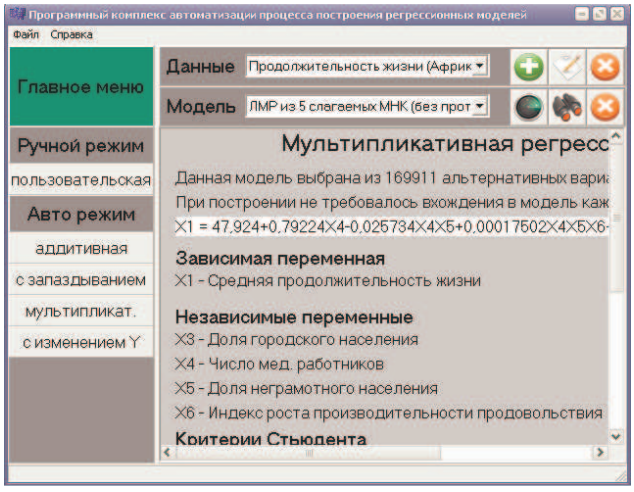

Рис. 1. Интерфейс ПК АППРМ.

В качестве примера решения задачи, демонстрирующего преимущества предложенного метода и программного комплекса, взяты классические данные работы выпарного аппарата на большом промышленном предприятии. Таблица с этими данными, состоящая из двадцати пяти наблюдений и десяти факторов, приведена в [1]. Далее дано подробное описание этих факторов.

1. Количество используемого пара в фунтах ежемесячно – x_1.

2. Количество активной жирной кислоты в фунтах, накопленное за месяц – x_2.

3. Количество готового глицерина в фунтах – x_3.

4. Средняя скорость ветра в милях в час – x_4.

5. Число календарных дней в месяце – x_5.

6. Число рабочих дней в месяце – x_6.

7. Число дней с температурой ниже 32 градусов по шкале Фаренгейта – x_7.

Средняя температура воздуха по шкале Фаренгейта – x_8.

Квадрат средней скорости ветра – x_9.

Число пусков – x_{10}.

По этим данным Н. Дрейпером и Г. Смитом было получено следующее линейное регрессионное уравнение:

$$x_1 = 9{,}1266 - 0{,}9724x_8 + 0{,}2029x_6. \qquad (6)$$

Критерии адекватности этой модели: $R = 0{,}849$, $F = 41{,}27$, $S = 0{,}438$, $E = 5{,}706\%$, $DW = 2{,}196$.

Согласно критериям, модель (6) достаточно значима и может использоваться, например, для получения прогнозных значений. Кроме того, разумны знаки коэффициентов регрессоров. Коэффициент при x_8 отрицателен. Действительно, чем выше температура воздуха, тем меньше нужно пара. С другой стороны, чем больше дней работает фабрика, тем больше расходуется пара, что отражается на знаке коэффициента x_6, который оказался положительным.

Перед тем, как начать моделирование с применением ПК АППРМ, были проанализированы исходные независимые переменные и выделены из них самые значимые по смыслу. Фактор x_5 – число календарных дней в месяце – логично исключить из рассмотрения, потому что он никак не влияет на производство пара. Действительно, не имеет значения, сколько дней в месяце – 28 или 31, важно только число рабочих дней, ведь по выходным никакого производства нет. Переменная x_7 – число дней с температурой ниже 32 градусов по шкале Фаренгейта – в исходной таблице содержит много нулевых значений, что не позволит работать в ПК АППРМ с некоторыми элементарными функциям, такими, например, как $\ln(x)$, $1/x$ и другие. Для устранения этой проблемы будем рассматривать переменную $(x_5 - x_7)$ – число дней с температурой не ниже 32 градусов, которая уже не содержит нулей. Фактор x_9 – квадрат средней скорости ветра – представляет собой преобразование x_4^2, и его тоже необходимо исключить из-за того, что такое преобразование будет производиться внутри комплекса, для чего достаточно указать только саму переменную x_4. Последняя переменная x_{10} – число пусков – также зависит от числа рабочих дней, и ее можно исключить из рассмотрения. В результате анализа выделен следующий набор из шести объясняющих факторов.

Количество активной жирной кислоты в фунтах, накопленное за месяц – x_2.

1. Количество готового глицерина в фунтах – x_3.

2. Средняя скорость ветра в милях в час – x_4.

3. Число рабочих дней в месяце – x_6.

4. Число дней с температурой не ниже 32 градусов по Фаренгейту – $(x_5 - x_7)$.

5. Средняя температура воздуха по Фаренгейту – x_8.

По смысловому содержанию разделим эти факторы на две группы:

1. Позитивные факторы, с ростом значений которых будет увеличиваться количество пара.

2. Негативные факторы, с ростом значений которых будет уменьшаться количество пара.

В результате анализа к позитивным факторам можно отнести x_2, x_3, x_4, x_6, а к негативным $(x_5 - x_7)$, x_8.

Теперь применим ПК АППРМ для построения регрессионной модели зависимости количества используемого пара x_1.

Сначала построим аддитивную модель, для чего выберем в комплексе следующие параметры поиска:

- зависимая переменная – x_1;
- независимые переменные – x_2, x_3, x_4, x_6, $x_5 - x_7$, x_8;
- метод оценивания – МНК;
- число регрессоров – 2;
- элементарные функции

$$-\frac{1}{x^2},\ \frac{1}{x},\ \frac{1}{\sqrt{x}},\ \sqrt{x},\ x,\ x^2,\ x^3,\ \ln x,\ 2^{0,4x},\ 3^{0,4x};$$

- критерии адекватности – R, F, S, E, DW.

В этом случае $p = 2$, $m = 6$, $l = 10$, следовательно, общее число альтернативных вариантов регрессий равно $r = C_{m \cdot l}^{p} = C_{60}^{2} = 1770$.

В результате работы комплекса из 1770 моделей были выбраны 596 моделей, удовлетворяющих смыслу задачи, а из них, исходя из заданных критериев адекватности, была выбрана лучшая модель:

$$x_1 = 23{,}1 + 0{,}004349 \cdot 2^{0,4x_6} - 3{,}876 \ln x_8. \qquad (7)$$

Критерии адекватности: $R = 0{,}908$, $F = 72{,}238$, $S = 0{,}267$, $E = 4{,}05\%$, $DW = 1{,}998$.

Очевидно, что полученная с помощью ПК АППРМ модель (7) оказалась лучше модели (6) по всем критериям адекватности. Также коэффициенты уравнения (7) удовлетворяют смыслу задачи, причем увеличение числа рабочих дней x_6 приводит к резкому возрастанию количества пара (показательная функция), а увеличение температуры воздуха x_8 – к более плавному убыванию количества пара (логарифмическая функция). Графики фактических и расчетных по модели (7) значений фактора x_1 приведены на рис. 2.

Построим теперь аддитивную модель с теми же параметрами поиска, но по методу наименьших модулей.

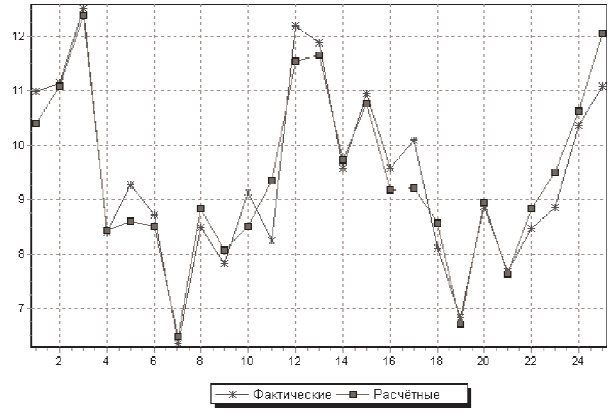

Рис. 2. График значений фактора x_1.

Из 1770 альтернатив было выбрано 565 уравнений, удовлетворяющих смыслу задачи, а из них выбрана одна лучшая:

$$x_1 = 24{,}02 + 0{,}004373 \cdot 2^{0{,}4x_6} - 4{,}11\ln x_8 . \qquad (8)$$

Критерии адекватности: $R = 0{,}905$, $F = 69{,}959$, $S = 0{,}275$, $E = 3{,}949\%$, $DW = 1{,}896$.

Получилось, что модель (8) содержит в своем составе те же регрессоры, что и уравнение (7). Это еще раз подтверждает правильность выбора формы связи для заданных параметров. Коэффициенты этой модели также удовлетворяют смыслу задачи, и по всем критериям она лучше линейной модели (6), но немного хуже модели (7).

Построим мультипликативную модель, в которой не требуется обязательного вхождения каждой независимой переменной. Для этого использованы те же самые параметры поиска, что и в предыдущем случае. Тогда общее число моделей равно $r = C^{p}_{2^m-1} = C^{2}_{63} = 1953$. В результате поиска из 1953 альтернатив 105 удовлетворяют смыслу задачи, а лучшая из них:

$$x_1 = 10{,}75 + 0{,}0244x_2x_6 - 0{,}077x_8 . \qquad (9)$$

Критерии адекватности: $R = 0{,}889$, $F = 58{,}607$, $S = 0{,}323$, $E = 4{,}39\%$, $DW = 1{,}728$.

Структура модели (9) отличается от структур полученных до этого регрессий тем, что в нее входит еще один фактор – количество жирной кислоты – x_2, и регрессор x_2x_6 говорит о том, что эти факторы оказывают совместное влияние на количество пара. Очевидно, что это уравнение удовлетворяет смыслу задачи: с ростом x_2 или x_6 количество пара увеличивается, а с возрастанием x_8, как и в предыдущих уравнениях, уменьшается. Модель (9) лучше линейной модели (6) по всем критериям, кроме критерия Дарбина-Уотсона, и в то же время она хуже по всем критериям, чем аддитивные модели (7) и (8).

Все три полученные с использованием ПК АППРМ регрессионные модели (7), (8) и (9) удовлетворяют смыслу задачи, и каждая из них оказалась по всем критериям адекватности, кроме модели (9), лучше классической линейной модели (6). Самой же лучшей из этих трех, очевидно, является аддитивная модель (7), полученная по МНК, и именно ее предлагается использовать на практике вместо зависимости (6). Как показывает практика, аддитивные модели не всегда оказываются лучше, чем мультипликативные. Все зависит от конкретной задачи и от исходных данных.

В заключение следует отметить, что реализованные в новой версии ПК АППРМ методы и алгоритмы не имеют аналогов в других программных продуктах, и, как видно из примера, применение этого комплекса при проведении регрессионного анализа оказывается весьма эффективным. Эта система может быть использована практически для любой предметной области, в которой имеются количественные статистические данные

Литература

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: пер. с англ. М.: Диалектика, 2007. 912 с.

2. Себер Дж. Линейный регрессионный анализ: пер. с англ. М.: Мир, 1980. 456 с.

3. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: исследование зависимостей. М.: Финансы и статистика, 1985. 487 с.

4. Афифи А., Эйзен С. Статистический анализ: подход с использованием ЭВМ. М.: Мир, 1982. 486 с.

5. Гайдышев И. Анализ и обработка данных: спец. справ. СПб: Питер, 2001. 752 с.

6. Орлов А.И. Прикладная статистика. М.: Экзамен, 2007. 672 с.

7. Клейнер Г.Б. Производственные функции. М.: Финансы и статистика, 1986. 239 с.

8. Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981. 304 с.

9. Miller A.J. Subset selection in regression // Chapman & Hall/CRC, 2002. P.247.

10. Ханк Д.Э., Уичерн Д.У., Райте А.Дж. Бизнес-

References

1. Dreyper N., Smith G. Applied regression analysis z: per. s angl. M.: Dialektika, 2007. 912 s.

2. Seber J. Linear regression analysis: per. s angl. M.: Mir, 1980. 456 s.

3. Aivazyan S.A., Enyukov I.S., Meshalkin L.D. Applied statistics: dependence study. M.: Finansy i statistika, 1985. 487 s.

4. Afifi A., Eizen S. Statistical analysis: computer-aided approach. M.: Mir, 1982. 486 s.

5. Gaidyshev I. Data processing analysis: a specialized reference book. Spb: Piter, 2001. 752 s.

6. Orlov A.I. Applied statistics. M.: Ekzamen, 2007. 672 s.

7. Kleyner G.B. Production functions. M.: Finansy i ststistika, 1986. 239 s.

8. Demidenko E.Z. Linear and nonlinear regressions. M.: Finansy i ststistika, 1981. 304 s.

9. Miller A.J. Subset selection in regression // Chapman & Hall/CRC, 2002. P.247.

10. Hank D.E., Whichern D.W., Rayte A.J. Business programming.

прогнозирование. М.: Вильямс, 2003. 656 с.

11. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. М.: Финансы и статистика, 1987. 239 с.

12. Edirisooriya G. Stepwise regression is a problem, not a solution // The Annual Meeting of the Mid-South Educational Research Association. Biloxi, 1995.

13. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. Киев: Наукова думка, 1981. 296 с.

14. Стрижов В.В. Методы индуктивного порождения регрессионных моделей. М.: ВЦ РАН, 2008. 54 с.

15. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. 3-е изд., перераб. и доп. М.: ИНФРА-М, 2003. 544 с.

16. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. 6-е изд., перераб. и доп. М.: Дело, 2004. 576 с.

17. Renfro C.G. A compendium of existing econometric software packages // Journal of Economic and Social Measurement. 2004. № 29. P. 359-409.

18. Хахаев И.А. Экономим на расчетах // Мир ПК. 2007. №7. С. 52-55.

19. Берндт Э. Практика эконометрики: классика и современность / пер. с англ. под ред. С.А. Айвазяна. М.: ЮНИТИ-ДАНА, 2005. 863 с.

20. Смирнова О.С. Программное обеспечение для статистического анализа // Заводская лаборатория. Диагностика материалов. 2008. Т. 74, № 5. С.68-74.

21. Базилевский М.П., Носков С.И. Анализ специализированного программного обеспечения для автоматизации «конкурса» регрессионных моделей // Информационные технологии и проблемы математического моделирования сложных систем. 2010, Вып.8. С.49-55.

22. Базилевский М.П., Носков С.И. Технология организации конкурса регрессионных моделей // Информационные технологии и проблемы математического моделирования сложных систем: сб. ст. Иркутск, 2009. Вып.7. С.77-84.

23. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: Облинформпечать, 1996. 320 с.

24. Макаров Н.М. Теория выбора и принятия решений. М.: Наука, 1982. 392 с.

25. Базилевский М.П., Носков С.И. Алгоритм построения линейно-мультипликативной регрессии // Современные технологии. Системный анализ. Моделирование. 2011. № 1 (29). С.90-94.

M.: Williams, 2003. 656 s.

11. Vuchkov I., Boyadzhiyeva l., Solakov E. Applied linear regression analysis. M.: Finansy i ststistika, 1987. 239 s.

12. Edirisooriya G. Stepwise regression is a problem, not a solution // The Annual Meeting of the Mid-South Educational Research Association. Biloxi, 1995.

13. Ivakhnenko A.G. Inductive method for complex system models self-organization. Kiev: Naukova dumka, 1981. 296 s.

14. Strizhov V.V. Methods for inductive generation of regression models. M.: VTs RAN, 2008. 54 s.

15. Tyurin Yu,N., Makarov A.A. Computer-aided data analysis. 3e izd., pererab. i dop. M.: INFRA-M, 2003. 544 s.

16. Magnus Ya.R., Katyshev P.K., Peresetsky A.A. Econometrics. The beginner's course. 6e izd., pererab. i dop. M.: Delo, 2004. 576 s.

17. Renfro C.G. A compendium of existing econometric software packages // Journal of Economic and Social Measurement. 2004. № 29. P. 359-409.

18. Khakhaev I.A. Let's save on computations // Mir PK.2007. №7. S. 52-55.

19. Berndt E. Econometrics practice: the classics and today / per. s angl. pod red. S.A. Aivazyana. M.: YuNITI-DANA, 2005. 863 s.

20. Smirnova O.S. Statistical analysis software // Zavodskaya laboratoriya. Diagnostika materialov. 2008. T. 74, № 5. S. 68-74.

21. Bazilevsky M.P., Noskov S.I. Specialized software analysis to automatize regression models «competition» // Informatsyonnye tekhnologii i problem matematicheskogo modelirovaniya slozhnykh system. 2010. Vyp. 8. S. 49-55/

22. Bazilevsky M.P., Noskov S.I. The techniques for organization of regression models competition // Informatsyonnye tekhnologii i problem matematicheskogo modelirovaniya slozhnykh system: sb. st. Irkutsk, 2009. Vyp. 7. S. 77-84.

23. Noskov S.I. Modeling technology for objects of unstable performance and data uncertainty. Irkutsk: Oblinformpechat', 1996. 320 s.

24. Makarov N.M. The theory of choice and decision theory. M.: Nauka, 1982. 320 s.

25. Bazilevsky M.P., Noskov S.I. The algorithm for linear-multiplicative regression construction // Sovremennye tekhnologii. Sistemny analiz. Modelirovaniye. 2011. № 1 (29). S. 90-94.

УДК 614.84

К вопросу разработки методики оценки уровня транспортной безопасности

С.И. Носков[1*], В.А. Протопопов[1]

[1]Иркутский государственный университет путей сообщения, Чернышевского 15, Иркутск, Россия.
Статья поступила 20.11 2011, принята 17.02.2012

В статье предлагается подход к разработке методики оценки уровня безопасности (опасности) объектов транспортной инфраструктуры, основанной на применении методов математического моделирования и предполагающей описание на формальном уровне уязвимости объектов и ущерба от ее реализации. При этом в качестве формальных конструкций, описывающих динамику указанных показателей, предлагается использовать широкий класс кусочных аппроксимирующих функций. Меру опасности (уязвимости) объекта традиционно предлагается рассчитывать в виде произведения вероятности нарушения транспортной безопасности (уязвимости) на ущерб от него. Ущерб принято оценивать как в абсолютных единицах (руб.), так и в относительных (процентах от общей стоимости основных производственных фондов). В работе для придания методике универсального характера использована вторая форма представления ущерба.

Ключевые слова: транспортная инфраструктура, уязвимость, безопасность, математическое моделирование.

* E-mail address: noskov_s@irgups.ru