

СТРУКТУРНАЯ И ПАРАМЕТРИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ЛИНЕЙНОЙ МНК-МОДЕЛИ БЕЗ РЕШЕНИЯ СИСТЕМЫ НОРМАЛЬНЫХ УРАВНЕНИЙ

В.В. Белов,

доктор технических наук, профессор кафедры «Вычислительная и прикладная математика» Рязанского государственного радиотехнического университета, Адрес: 390005, Рязань, ул. Гагарина, 59/1, кафедра «Вычислительная и прикладная математика» Белов В.В. Тел: 8-903-835-54-98. E-mail: compvv@mail.ryazan.ru

Предлагается способ вычисления значений линейной МНК-модели, в частности MLR (Multiple Linear Regression), без оценивания её параметров. Предлагаемый способ может использоваться в алгоритмах поиска наилучшего в некотором смысле линейного описания процесса, представленного дискретными значениями (временным рядом). Предлагается рекуррентная схема вычисления параметров МНК-модели, альтернативная решению системы нормальных уравнений Гаусса.

Ключевые слова: Линейная МНК-модель. Последовательное введение переменных. Вычисление значений. Оценка параметров. Рекурсивная схема.

Предварительные замечания

Результаты, приведённые в данной статье, получены в процессе выполнения работ по заказам Министерства труда и социального развития России, а также Федеральной службы государственной статистики. Главные задачи работ состояли в получении прогнозных значений для заданных групп показателей производственного травматизма в РФ и занятости населения в экономике страны. Необходимые статистические данные были предоставлены Федеральной службой государственной статистики. Обусловливалось обязательное построение варианта модели в виде линейной множественной регрессии (MLR-описания) с регрессорами, входящими в сценарные условия прогнозирования. Допускалось использование альтернативных методов моделирования для дополнительной оценки надежности полученного прогноза. Поскольку крайне желательно находить MLR-описание с наименьшим количеством параметров (это, в частности, способствует минимизации дисперсии коэффициентов модели), в практике

реальных приложений регрессионного анализа часто приходится решать задачу структурной идентификации линейной модели – поиска «наилучшей» в некотором смысле регрессии. В процессе решения этой задачи происходит последовательное добавление регрессоров в модель в разных сочетаниях до достижения требуемой погрешности аппроксимации.

Длительные упражнения в решении задачи последовательного синтеза наилучшей линейной модели и оценки качества частичных описаний привели к возникновению идеи вычисления векторов значений MLR-описания без предварительного оценивания его параметров, т.е. без осуществления операции параметрической идентификации. Такой приём позволяет ускорить процесс поиска: новый регрессор добавляется без повторного решения системы нормальных уравнений Гаусса. Кроме того, происходит, хотя и несущественное, повышение точности вычислений. Платой за эти преимущества являются ухудшение пространственных характеристик программного приложения – увеличиваются затраты на память.

Сопутствующим результатом явился альтернативный алгоритм параметрической идентификации – вычисления параметров линейной модели без решения системы нормальных уравнений. Однако он представляет, видимо, чисто теоретический интерес – как средство описания вариативности процедур оценки параметров линейных описаний. В плане утилитарности вряд ли можно указать условия целесообразности его применения.

История вопроса

Задача расширения линейной модели, параметры которой оцениваются методом наименьших квадратов, впервые рассмотрена в [1]. Однако В.М. Кохран ограничился рассмотрением случая добавления в модель одной переменной. На случай нескольких переменных подход В.М. Кохрана распространил М.Г. Квинауил [2]. Указанные результаты описаны в [3] и доступны на русском языке в переводе [4]. Общим началом подходов В.М. Кохрана и М.Г. Квинауила является то, что они предлагают схему добавления новых регрессоров, меняющую коэффициенты предшествующих описаний: добавление нового слагаемого $(a_p \cdot x_p)$ в MLR-описание сопровождается пересчетом всех коэффициентов a_1, a_2, \dots, a_{p-1} предыдущего варианта модели.

Последовательное расширение описания, предлагаемое в настоящей статье, отличается тем, что параметры модели не оцениваются вовсе:

1) коэффициент нового добавляемого в модель регрессора не вычисляется;

2) значения коэффициентов «старых» регрессоров не пересчитываются.

Предлагаемый способ целесообразен в тех случаях, когда конкретика значений коэффициентов не важна. Это, прежде всего, как уже указывалось, – задачи поиска наилучших линейных моделей с минимальным количеством параметров.

Классическое решение задачи последовательного введения регрессоров

При введении дополнительных регрессоров по Кохрану и Квинауилу предполагается, что подобрана модель регрессии

$$E(Y) = X\beta, \quad D(Y) = \sigma^2 I_n,$$

где $Y = (y_1, y_2, \dots, y_n)^T$ – временной ряд, т.е. вектор последовательных равноотстоящих по времени

значений некоторого процесса (результативного признака); X – матрица регрессоров (аргументов или факторных признаков) размером $n \times p$ ранга p ; β – вектор коэффициентов регрессии, состоящий из p элементов; I_n – единичная матрица порядка n ; σ – дисперсия значений временного ряда Y . Оценка $\hat{\beta}$ вектора β находится методом наименьших квадратов: $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Рассматривается задача добавления новых регрессоров таким образом, чтобы новая модель имела вид:

$$G: E(Y) = X\beta_G + Z\gamma_G = (X \quad Z) \begin{pmatrix} \beta_G \\ \gamma_G \end{pmatrix} = W\delta_G,$$

где буква G идентифицирует новую модель;

Z – матрица дополнительных регрессоров размером $n \times t$ ранга t ;

β_G – вектор новых коэффициентов при «старых» регрессорах, объединённых в матрицу X ;

γ_G – вектор коэффициентов при новых регрессорах, объединённых в матрицу Z , состоящий из t элементов;

$W = (X \quad Z)$ – матрица объединённых регрессоров, т.е. матрица всех регрессоров модели G , имеющая размер $n \times (p + t)$ и ранг $p + t$;

$\delta_G = (\beta_G \quad \gamma_G)^T$ – вектор всех коэффициентов модели G , состоящий из $p + t$ элементов.

Указанные предположения означают, что имеет место стандартная ситуация: длина n временного ряда Y больше числа параметров модели, т.е. $n > p + t$ и все столбцы матрицы W (все регрессоры) линейно не зависимы.

Естественно, вычислить оценку $\hat{\delta}_G$ вектора δ_G и её дисперсионную матрицу $D(\hat{\delta})$ можно следующим:

$$\hat{\delta}_G = (W^T W)^{-1} W^T Y, \quad D(\hat{\delta}) = \sigma^2 (W^T W)^{-1}.$$

Однако можно сократить объём вычислений, используя результат обращения матрицы $(X^T X)$, полученный ранее при вычислении оценки $\hat{\beta}$ вектора коэффициентов β .

Суть метода Кохрана-Квинауила выражается следующей теоремой [4, с. 69].

Теорема Себера

Пусть $R = I_n - X(X^T X)^{-1} X^T$,

$$R_G = I_n - W(W^T W)^{-1} W^T,$$

$$L = (X^T X)^{-1} X^T Z,$$

$$M = [Z^T R Z]^{-1} \text{ и } \hat{\delta}_G = \begin{pmatrix} \hat{\beta}_G \\ \hat{\gamma}_G \end{pmatrix}.$$

Тогда

$$\begin{aligned} \text{(i)} \quad \hat{\beta}_G &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{Z} \hat{\gamma}_G) = \hat{\beta} - \mathbf{L} \hat{\gamma}_G. \\ \text{(ii)} \quad \hat{\gamma}_G &= (\mathbf{Z}^T \mathbf{R} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{R} \mathbf{Y}. \\ \text{(iii)} \quad \mathbf{Y}^T \mathbf{R}_G \mathbf{Y} &= (\mathbf{Y} - \mathbf{Z} \hat{\gamma}_G)^T \mathbf{R} (\mathbf{Y} - \mathbf{Z} \hat{\gamma}_G). \\ \text{(iv)} \quad \mathbf{Y}^T \mathbf{R}_G \mathbf{Y} &= \mathbf{Y}^T \mathbf{R} \mathbf{Y} - \hat{\gamma}_G^T \mathbf{Z}^T \mathbf{R} \mathbf{Y} \\ \text{(v)} \quad D(\hat{\delta}_G) &= \sigma^2 \begin{pmatrix} (\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{L} \mathbf{M} \mathbf{L}^T & -\mathbf{L} \mathbf{M} \\ -\mathbf{M} \mathbf{L}^T & \mathbf{M} \end{pmatrix}. \end{aligned}$$

Замечания:

1) пункт (ii) теоремы определяет алгоритм вычисления коэффициентов при новых переменных, добавляемых в модель; если вместо нескольких переменных добавить одну переменную, то матрица \mathbf{Z} трансформируется в вектор \mathbf{z} , одновременно матрицы $\mathbf{Z}^T \mathbf{R} \mathbf{Z}$ и $\mathbf{Z}^T \mathbf{R} \mathbf{Y}$ трансформируются в скаляры, формула коэффициента при добавляемой переменной принимает вид:

$$\hat{\gamma}_G = \frac{\mathbf{z}^T \mathbf{R} \mathbf{Y}}{\mathbf{z}^T \mathbf{R} \mathbf{z}},$$

и это уже не вектор, а скаляр;

2) пункт (i) определяет алгоритм коррекции вектора $\hat{\beta}$ коэффициентов при переменных, уже входивших в модель; заметим, что он предполагает использование вектора $\hat{\gamma}_G$ коэффициентов при добавляемых переменных, определяемого в следующем пункте, и, главное, — формула для матрицы \mathbf{L} предполагает вычисление обратной матрицы $(\mathbf{X}^T \mathbf{X})^{-1}$;

3) семантически матрица $\mathbf{L} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Z}$ представляет собой совокупность векторов модельных значений — j -й столбец этой матрицы представляет собой результат объяснения j -й добавляемой переменной всеми предыдущими переменными;

4) пункты (iii) и (iv) определяют две эквивалентные по значениям формулы вычисления остаточной суммы квадратов $\mathbf{Y}^T \mathbf{R}_G \mathbf{Y}$ для итоговой модели через матрицы \mathbf{R} и \mathbf{Z} , формируемые старыми и новыми переменными.

Предлагаемое решение задачи последовательного введения регрессоров без параметрической идентификации

Теорема 1

Пусть $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ — вектор вещественных чисел, условно называемый результативным признаком;

$$\begin{aligned} \mathbf{x}_1 &= (x_{1,1}, x_{2,1}, \dots, x_{n,1})^T, \\ \mathbf{x}_2 &= (x_{1,2}, x_{2,2}, \dots, x_{n,2})^T, \\ \mathbf{x}_p &= (x_{1,p}, x_{2,p}, \dots, x_{n,p})^T \end{aligned}$$

векторы вещественных чисел, условно называемые объясняющими переменными или факторными признаками; $\mathbf{J}_n = [1]_1^n$ — вектор единиц, состоящий из n элементов; \mathbf{I}_n — единичная матрица размером $n \times n$, $\mathbf{X} = [\mathbf{x}_0 \mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_p]$ — матрица объясняющих переменных, причём $\mathbf{x}_0 = \mathbf{J}_n$ и $\text{rank}(\mathbf{X}) = p + 1$; $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$, — вектор значений линейной модели зависимости результативного признака от объясняющих переменных, т.е. $\hat{\mathbf{y}} = \mathbf{X} \cdot \mathbf{a}$, где $\mathbf{a} = (a_0, a_1, \dots, a_p)$ — вектор параметров (коэффициентов) модели, вычисляемых методом наименьших квадратов: $\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$; $\mathbf{P} = \mathbf{X} \cdot (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ — матрица, порождающая значения модели $\hat{\mathbf{y}} = \mathbf{P} \cdot \mathbf{y}$.

Тогда:

1) проекционная матрица $\mathbf{P} = \mathbf{X} \cdot (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, порождающая значения $\hat{\mathbf{y}}$ линейной модели, построенной с использованием факторных признаков x_1, x_2, \dots, x_p , эквивалентной по значениям описанию, параметры которого оцениваются методом наименьших квадратов, может быть получена по рекуррентной схеме: $\mathbf{P} = \mathbf{P}_p$, где \mathbf{P}_p — финальное значение в последовательности рекуррентных вычислений

$$\begin{aligned} \mathbf{P}_j &= \mathbf{P}_{j-1} + \frac{\mathbf{R}_{j-1} \cdot \mathbf{x}_j \cdot \mathbf{x}_j^T \cdot \mathbf{R}_{j-1}}{\mathbf{x}_j^T \cdot \mathbf{R}_{j-1} \cdot \mathbf{x}_j}, \quad j = \overline{1, p}; \\ \mathbf{R}_{j-1} &= \mathbf{I}_n - \mathbf{P}_{j-1}; \\ \mathbf{P}_0 &= \frac{\mathbf{J}_n \cdot \mathbf{J}_n^T}{\mathbf{J}_n^T \cdot \mathbf{J}_n} = \frac{\mathbf{J}_n \cdot \mathbf{J}_n^T}{n}; \end{aligned}$$

2) остаточная сумма квадратов $RSS = \mathbf{y}^T \cdot (\mathbf{I}_n - \mathbf{P}_p) \cdot \mathbf{y}$.

Доказательство

Запишем формулу (3.37) [4, с. 72], описывающую значения модели с ортогональной структурой, для частного случая, когда матрица \mathbf{Z} вырождается в вектор \mathbf{x}_j в новых обозначениях:

$$\hat{\mathbf{y}}^{[j]} = \mathbf{X}_{j-1} \cdot \mathbf{a}_{j-1}^{[j-1]} + \mathbf{R}_{j-1} \cdot \mathbf{x}_j \cdot a_j^{[j]}. \quad (*)$$

В соответствии с пунктом (ii) теоремы Себера (см. выше) для рассматриваемого частного случая и используемых обозначений имеем:

$$a_j^{[j]} = \frac{\mathbf{x}_j^T \cdot \mathbf{R}_{j-1} \cdot \mathbf{y}}{\mathbf{x}_j^T \cdot \mathbf{R}_{j-1} \cdot \mathbf{x}_j}.$$

Кроме того, учтём, что

$$\mathbf{X}_{j-1} \cdot \mathbf{a}_{j-1}^{[j-1]} = \hat{\mathbf{y}}^{[j-1]}.$$

Подставляя указанные выражения в (*), получим:

$$\hat{\mathbf{y}}^{[j]} = \hat{\mathbf{y}}^{[j-1]} + \mathbf{R}_{j-1} \cdot \mathbf{x}_j \cdot \frac{\mathbf{x}_j^T \cdot \mathbf{R}_{j-1} \cdot \mathbf{y}}{\mathbf{x}_j^T \cdot \mathbf{R}_{j-1} \cdot \mathbf{x}_j}.$$

Далее с учётом того, что

$$\hat{y}^{[j]} = P_j \cdot y \quad \text{и} \quad \hat{y}^{[j-1]} = P_{j-1} \cdot y$$

имеем

$$P_j \cdot y = P_{j-1} \cdot y + \frac{R_{j-1} \cdot x_j \cdot x_j^T \cdot R_{j-1}}{x_j^T \cdot R_{j-1} \cdot x_j} \cdot y.$$

Откуда и следует, что матрица P_j может быть вычислена рекуррентно:

$$P_j = P_{j-1} + \frac{R_{j-1} \cdot x_j \cdot x_j^T \cdot R_{j-1}}{x_j^T \cdot R_{j-1} \cdot x_j}.$$

Терминальная ветвь рекурсии для случая $j = 1$ определяется значением P_0 . По определению

$$P_0 = x_0 \cdot (x_0^T \cdot x_0)^{-1} \cdot x_0^T.$$

Поскольку $x_0 = J_n$, а произведение $J_n^T \cdot J_n$ скалярно и равно n , имеем:

$$P_0 = \frac{J_n \cdot J_n^T}{n}.$$

Остаточная сумма квадратов RSS по определению равна $y^T \cdot (I_n - P) \cdot y$. Равенство $P = P_p$ определяет возможность вычисления остаточной суммы по формуле $RSS = y^T \cdot (I_n - P_p) \cdot y$.

Алгоритм вычислений

В соответствии с указанной теоремой алгоритм вычисления вектора значений \hat{y} линейной модели без операции её параметрической идентификации имеет вид:

3) вспомогательные величины:

(1) $J_n = [1]_1^n$ – вектор единиц, состоящий из n элементов;

(2) $I_n = \text{identity}(n)$ – единичная матрица размером $n \times n$;

4) начальное значение:

$$P_0 = \frac{J_n \cdot J_n^T}{J_n^T \cdot J_n} = \frac{J_n \cdot J_n^T}{n}$$

квадратная матрица размером $n \times n$, все элементы которой равны $1/n$, – начальное значение матрицы, порождающей линейную модель, построенную по первому столбцу $x_0 = J_n$ матрицы объясняющих переменных X_p ;

5) рекуррентные вычисления для $j = \overline{1, p}$:

(1) $R_{j-1} = I_n - P_{j-1}$ – вспомогательная матрица для упрощения записи формул;

(2) если $R_{j-1} \cdot x_j \neq [0]_1^n$, где $[0]_1^n$ – вектор, состоящий из одних нулей, то вычисляется

$$P_j = P_{j-1} + \frac{R_{j-1} \cdot x_j \cdot x_j^T \cdot R_{j-1}}{x_j^T \cdot R_{j-1} \cdot x_j} -$$

очередное значение порождающей матрицы P ;

(3) иначе, если $R_{j-1} \cdot x_j = [0]_1^n$, то переменная x_j исключается из списка потенциальных аргументов модели и осуществляется переход к следующему значению j ;

6) вычисляется вектор $\hat{y}(X) = P_p \cdot y$ значений модели, построенной по матрице объясняющих переменных $X = [x_0 \ x_1 \ x_2 \ \dots \ x_p]$.

Замечания:

1) вычисление значений модели по указанному алгоритму не требует выполнения операции обращения матрицы;

2) на каждом шаге рекуррентных вычислений могут быть получены значения частных описаний, построенных по части объясняющих переменных: $y^{[j]} = \hat{y}(x_0 \ x_1 \ x_2 \ \dots \ x_p) = P_j \cdot y$;

3) условие $R_{j-1} \cdot x_j \neq [0]_{n \times 1}$ исключает переменные, векторы значений которых коллинеарны с векторами уже включёнными в модель; произведение $R_{j-1} \cdot x_j$ представляет собой вектор ошибок объяснения вектора x_j векторами $x_0 \ x_1 \ x_2 \ \dots \ x_{j-1}$.

Вычислительный эксперимент

Проверка правильности алгоритма была осуществлена путём генерации нескольких векторов случайных чисел, равномерно распределённых от нуля до десяти. Один из векторов интерпретировался как значения результирующего признака $y = (y_1, y_2, \dots, y_n)^T$, а остальные – как значения факторных признаков $x_1 = (x_{1,1}, x_{2,1}, \dots, x_{n,1})^T$, $x_2 = (x_{1,2}, x_{2,2}, \dots, x_{n,2})^T$, ..., $x_p = (x_{1,p}, x_{2,p}, \dots, x_{n,p})^T$. Матрица, порождающая значения модели находилась двумя способами – рекуррентно P_p и путём прямых вычислений $P = X \cdot (X^T \cdot X)^{-1} \cdot X^T$. Затем находились векторы значений модели $\hat{y}^{[p]} = P_p \cdot y$ и $\hat{y} = P_p \cdot y$.

Длины векторов, участвовавших в вычислениях выбрались равными: 5; 10; 30; 100 и 500. Количество факторных признаков: 4; 7 и 15.

Расчёты выполнялись с использованием восьми байтов для представления вещественных чисел. Результаты показали, что во всех случаях абсолютная величина разности элементов векторов $\hat{y}^{[p]}$ и \hat{y} мала:

$$\max(\overline{|\hat{y}^{[p]} - \hat{y}|}) < 10^{-14},$$

где стрелка над выражением символизирует операцию векторизации, в данном случае – формирования вектора абсолютных значений разностей

соответствующих элементов матриц и векторов. Таким образом, различие результатов вычисления проекционных матриц и значений модели классическим и предлагаемым способами имеет порядок погрешности внутримашинного представления вещественных данных.

Теорема 2

Пусть $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ – вектор вещественных чисел, условно называемый результативным признаком;

$$\begin{aligned} \mathbf{x}_1 &= (x_{1,1}, x_{2,1}, \dots, x_{n,1})^T, \\ \mathbf{x}_2 &= (x_{1,2}, x_{2,2}, \dots, x_{n,2})^T, \\ \mathbf{x}_p &= (x_{1,p}, x_{2,p}, \dots, x_{n,p})^T \end{aligned}$$

векторы вещественных чисел, условно называемые объясняющими переменными или факторными признаками; $\mathbf{J}_n = [1]_1^n$ – вектор единиц, состоящий из n элементов; \mathbf{I}_n – единичная матрица размером $n \times n$, $\mathbf{X} = [\mathbf{x}_0 \ \mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_p]$ – матрица объясняющих переменных, причём $\mathbf{x}_0 = \mathbf{J}_n$ и $\text{rang}(\mathbf{X}) = p + 1$; $\mathbf{a} = (a_0, a_1, \dots, a_p)$ – вектор параметров (коэффициентов) модели, вычисляемых методом наименьших квадратов: $\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \cdot \mathbf{X}^T \mathbf{y}$.

Тогда вектор параметров \mathbf{a} линейной модели может быть получен альтернативно по рекуррентной схеме:

$$\begin{aligned} a_p &= \frac{\mathbf{x}_p^T \cdot \mathbf{R}_{p-1} \cdot \mathbf{y}}{\mathbf{x}_p^T \cdot \mathbf{R}_{p-1} \cdot \mathbf{x}_p}; \\ a_k &= \frac{\mathbf{x}_k^T \cdot \mathbf{R}_{k-1} \cdot (\mathbf{y} - \sum_{j=k+1}^p a_j \mathbf{x}_j)}{\mathbf{x}_k^T \cdot \mathbf{R}_{k-1} \cdot \mathbf{x}_k}, \\ k &= p-1, p-2, \dots, 1; \\ a_0 &= \frac{\mathbf{J}_n^T \cdot (\mathbf{y} - \sum_{k=1}^p a_k \mathbf{x}_k)}{n}. \end{aligned}$$

Доказательство

Утверждение 1. Обозначим $\mathbf{X}_k = [\mathbf{x}_0 \ \mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_k]$, $k \leq p$. При этом $\mathbf{X} = [\mathbf{x}_0 \ \mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_p] = \mathbf{X}_p$. Согласно уравнениям системы $\mathbf{X}^T \cdot \mathbf{X} \cdot \mathbf{a} = \mathbf{X}^T \cdot \mathbf{y}$ нормальных уравнений Гаусса равенство справедливо для всех $j = \overline{0; p}$, поскольку $\mathbf{X}^T \cdot (\mathbf{y} - \mathbf{X} \cdot \mathbf{a}) = [0]_1^{p+1}$ и $\text{rang}(\mathbf{X}) = p+1$. Вследствие этого равенства $\mathbf{X}_k^T \cdot \mathbf{e} = [0]_1^{p+1}$ справедливы для всех $k = \overline{0; p}$.

Утверждение 2. Если в качестве объясняемой переменной использовать одну из объясняющих переменных \mathbf{x}_j , где $j \leq k$, с числовым коэффициентом a , то вектор $\mathbf{b} = (b_0, b_1, \dots, b_k)^T$, представляющий собой решение соответствующей системы нормальных уравнений Гаусса $\mathbf{X}_k^T \cdot \mathbf{X}_k \cdot \mathbf{b} = \mathbf{X}_k^T \cdot \mathbf{x}_j a$, таков: $b_j = a$; $b_i = 0$ при $0 \leq i < j$ и $j < i \leq k$, т.е. $b_i = 0$ при $i \neq j$.

Утверждение 3. Рассмотрим систему алгебраических уравнений

$$\mathbf{X}_k^T \cdot \mathbf{X}_k \cdot \mathbf{b} = \mathbf{X}_k^T \cdot (\mathbf{y} - \sum_{j=k+1}^p a_j \mathbf{x}_j).$$

относительно вектора \mathbf{b} . Заметим, что

$$\mathbf{y} - \sum_{j=k+1}^p a_j \mathbf{x}_j = \sum_{j=0}^k a_j \mathbf{x}_j + \mathbf{e}.$$

Рассматриваемая система принимает вид

$$\mathbf{X}_k^T \cdot \mathbf{X}_k \cdot \mathbf{b} = \mathbf{X}_k^T \cdot (\sum_{j=0}^k a_j \mathbf{x}_j + \mathbf{e})$$

и с учётом утверждений 1 и 2 имеем в качестве решения $\mathbf{b} = (a_0, a_1, \dots, a_k)$, т.е. вектор \mathbf{b} является частью вектора \mathbf{a} – состоит из первых $k + 1$ элементов этого вектора.

Утверждение 4. Формула для вычисления a_p определяет значение последнего (с наибольшим номером) коэффициента линейной модели согласно пункту (ii) теоремы Себера для частного случая $\mathbf{Z} = \mathbf{x}_p$.

Утверждение 5. Формула

$$\frac{\mathbf{x}_k^T \cdot \mathbf{R}_{k-1} \cdot (\mathbf{y} - \sum_{j=k+1}^p a_j \mathbf{x}_j)}{\mathbf{x}_k^T \cdot \mathbf{R}_{k-1} \cdot \mathbf{x}_k}$$

определяет значение последнего (k -го) коэффициента линейной модели для объясняемого вектора

$$\mathbf{y} - \sum_{j=k+1}^p a_j \mathbf{x}_j.$$

Согласно утверждению 3 этот коэффициент совпадает по значению с искомым коэффициентом a_k .

Утверждение 6. Заметим предварительно, что среднее значение \bar{v} произвольного вектора \mathbf{v} равно

$$\bar{v} = \frac{\mathbf{J}_n^T \cdot \mathbf{v}}{\mathbf{J}_n^T \cdot \mathbf{J}_n}.$$

При этом $\mathbf{J}_n^T \cdot \mathbf{J}_n = n$. Формула для вычисления a_0 получается как результат объяснения остатка

$$\mathbf{y} - \sum_{j=1}^p a_j \mathbf{x}_j$$

вектора \mathbf{y} вектором $\mathbf{x}_0 = \mathbf{J}$, т.е. это среднее значение вектора

$$\mathbf{y} - \sum_{j=1}^p a_j \mathbf{x}_j.$$

Пример практического использования предлагаемых решений

Предлагаемый способ последовательного введения регрессоров без параметрической идентификации был использован, в частности, для решения задачи прогноза значений группы показателей

производственного травматизма, занятости населения, реально располагаемых денежных доходов и заработной платы в РФ. В качестве исходных данных использованы временные ряды Государственного комитета по статистике РФ.

Прогнозируемые показатели для упрощения ссылок пронумерованы следующим образом:

T_1 – «Число пострадавших с утратой трудоспособности на 1 рабочий день и более (в том числе со смертельным исходом), тыс. чел.»;

T_2 – «Число пострадавших со смертельным исходом, чел.»;

T_3 – «Число пострадавших с утратой трудоспособности на 1 рабочий день и более (в том числе со смертельным исходом), в расчете на 1000 работающих»;

T_4 – «Число пострадавших со смертельным исходом в расчете на 1000 работающих»;

T_5 – «Число человеко-дней нетрудоспособности на 1 рабочий день и более, временная нетрудоспособность которых закончилась в отчетном году, в расчете на 1000 человек».

Θ_1 – «Численность занятого в экономике населения, млн чел.»;

Θ_2 – «Общая численность безработных, млн чел.»;

Θ_3 – «Реально располагаемые денежные доходы населения в % к предыдущему году»;

Θ_4 – «Номинальная начисленная среднемесячная заработная плата на 1 работника, руб.».

Показатели $T_1 - T_5$, Θ_1 , Θ_2 , Θ_4 – являются абсолютными (не относительными) и имеют натуральное выражение; Θ_3 – цепной годовой индекс, измеряемый в процентах.

Набор факторных переменных состоял из 12 показателей, каждый из которых в процессе предварительного корреляционного анализа учитывался с лагами 0, 1, 2 и 3 года. Таким образом, общее число потенциальных регрессоров составило 48. В результате поиска «лучших» описаний для каждой модели был определен адекватный набор регрессоров. Множество всех адекватных регрессоров финальных моделей включает:

1) ВВП – валовой внутренний продукт в сопоставимых ценах; используется в 5 моделях;

2) ВВП₂ – ВВП с лагом в 2 года; используется в двух моделях;

3) ПотрЦен – индекс потребительских цен; используется в двух моделях;

4) ПотрЦен-2 – ПотрЦен с лагом в 2 года; используется в одной модели;

5) ИнвОК – инвестиции в основной капитал в сопоставимых ценах; используется в трех моделях;

6) ИнвОК₂ – ИнвОК с лагом в 2 года; используется в трех моделях;

7) ОПП – объем промышленного производства в сопоставимых ценах; используется в одной модели;

8) ОПП₁ – ОПП с лагом в 1 год; используется в одной модели;

9) ЧисЗан – численность занятого в экономике населения; используется в трех моделях;

10) ЧисЗан₁ – ЧисЗан с лагом в 1 год; используется в одной модели;

11) ЧисЗан₂ – ЧисЗан с лагом в 2 года; используется в одной модели.

В данных Государственного комитета по статистике РФ значения факторных показателей представлены в абсолютном выражении и в виде цепных индексов. В то же время известно, что факторные показатели целесообразно представлять в виде базисных индексов, так как проблемные показатели более коррелированы с базисными индексами, нежели с цепными. Кроме этого, базирование отображает данные в окрестность интервала [0; 100], что приводит к одному порядку значений коэффициентов регрессии.

Исходные статистические данные для показателей производственного травматизма начинаются с 1993 г., данные по социально-экономическим показателям – с 1991 г. Этот факт обусловил то, что в качестве информационной платформы для построения моделей использовались показатели социально-трудовой сферы, начиная с 1993 г. В качестве базы использованы данные за 1991 г. Базисные индексы этого года приняты равными 100 %. Расчёт базисных индексов для k -го года через цепные осуществляется с помощью соотношения:

$$J_X(k) = J_X(k-1) \cdot j_X(k) / 100,$$

где $J_X(k)$, $j_X(k)$ – соответственно базисный и цепной годовой индексы фактора X .

В качестве сценарных условий для прогнозирования использованы «Основные показатели прогноза социально-экономического развития Российской Федерации до 2011 года» от Минэкономразвития РФ.

Окончательная модель для показателя T_1 является четырехпараметрической с тремя факторами без лагов:

$$T_1(t) = a_0 + a_1 \cdot J_{\text{ВВП}}(t) + a_2 \cdot J_{\text{ИнвОК}}(t) + a_3 \cdot J_{\text{ЧисЗан}}(t),$$

где t – номер года, $t \geq 1993$; $a_0 = -598,11$; $a_1 = -8,94$; $a_2 = 7,81$; $a_3 = 12,86$.

Окончательная модель для показателя T_2 содержит два фактора без лагов:

$$T_2(t) = a_0 + a_1 \cdot J_{\text{ВВП}}(t) + a_2 \cdot J_{\text{ИнвОК}}(t),$$

где t – номер года, $t \geq 1993$; $a_0 = 12341,26$;
 $a_1 = -241,06$; $a_2 = 270,71$.

Окончательная модель для показателя T_3 содержит также два фактора без лагов:

$$T_3(t) = a_0 + a_1 \cdot J_{\text{ИнвОК}}(t) + a_2 \cdot J_{\text{чисЗан}}(t),$$

где t – номер года, $t \geq 1993$; $a_0 = -15,04$; $a_1 = -0,0653$;
 $a_2 = 0,257$.

Окончательная модель для показателя T_4 имеет три фактора:

$$T_4(t) = a_0 + a_1 \cdot J_{\text{ВВП}}(t) + a_2 \cdot J_{\text{ОПП}}(t) + a_3 \cdot J_{\text{чисЗан}}(t-1),$$

где t – номер года, $t \geq 1994$; $a_0 = 0,476$; $a_1 = 0,00776$;
 $a_2 = -0,00589$; $a_3 = 0,00567$. Индекс численности занятых имеет лаг в 1 год.

Для показателя T_5 финальная модель содержит два фактора:

$$T_5(t) = a_0 + a_1 \cdot J_{\text{ИнвОК}}(t-2) + a_2 \cdot J_{\text{чисЗан}}(t),$$

где t – номер года, $t \geq 1995$; $a_0 = 12,24$; $a_1 = -0,089$;
 $a_2 = 0,212$. Индекс инвестиций в основной капитал имеет лаг в 2 года.

Для показателя Θ_1 модель содержит два фактора без лагов:

$$\Theta_1(t) = a_0 + a_1 \cdot J_{\text{ВВП}}(t) + a_2 \cdot J_{\text{ПотрЦен}}(t),$$

где t – номер года, $t \geq 1993$; $a_0 = 49,68$; $a_1 = -0,2748$;
 $a_2 = -0,0006274$.

Для показателя Θ_2 модель содержит три фактора с лагами в 2 года и 1 год:

$$\Theta_2(t) = a_0 + a_1 \cdot J_{\text{ВВП}}(t-2) + a_2 \cdot J_{\text{ИнвОК}}(t-2) + a_3 \cdot J_{\text{ОПП}}(t-1),$$

где t – номер года, $t \geq 1995$; $a_0 = 11,08$; $a_1 = 0,2884$;
 $a_2 = -0,2589$; $a_3 = -0,2466$.

Для показателя Θ_3 модель содержит три фактора:

$$\Theta_3(t) = a_0 + a_1 \cdot J_{\text{ВВП}}(t) + a_2 \cdot J_{\text{ПотрЦен}}(t) + a_3 \cdot J_{\text{чисЗан}}(t-2),$$

где t – номер года, $t \geq 1995$; $a_0 = 401,84$; $a_1 = 3,63$;
 $a_2 = -0,00974$; $a_3 = -5,563$. Индекс численности занятых имеет лаг в 2 года.

Для показателя Θ_4 модель содержит три фактора с лагами в 2 года:

$$\Theta_4(t) = a_0 + a_1 \cdot J_{\text{ВВП}}(t-2) + a_2 \cdot J_{\text{ПотрЦен}}(t-2) + a_3 \cdot J_{\text{ИнвОК}}(t-2),$$

где t – номер года, $t \geq 1995$; $a_0 = -2899,9$; $a_1 = 60,554$;
 $a_2 = 0,51199$; $a_3 = -26,595$.

Все полученные модели имеют хорошие показатели качества:

1) они значимы в целом по критерию Фишера на уровне значимости $\alpha = 0,01$;

2) все коэффициенты отличны от нуля по t -критериям на том же уровне значимости, таким образом, все модели являются приемлемыми по критерию минимальной значимой разности (least significant difference);

3) показатель прогностической силы по Дрейперу и Смиуту [5]

$$\frac{F}{F_{k_1, k_2}^{0,01}}$$

всех полученных описаний существенно (в два раза и более) превышает пороговый уровень 4, т.е. модели пригодны для решения задачи прогнозирования.

На основании изложенного можно утверждать, что прогнозы, получаемые по полученным описаниям имеют надежность, совпадающую с надежностью сценарных условий. Наиболее существенная качественная особенность полученных результатов прогноза состоит в следующем:

1) для показателей социально-экономического развития РФ Θ_1 , Θ_3 , Θ_4 полученные прогнозные значения количественно и качественно близки к прогнозу Минэкономразвития России, что свидетельствует о том, что оба прогноза отражают одни и те же взаимосвязи между показателями и факторами;

2) для показателя Θ_2 («Общая численность безработных, млн чел.»); полученные прогнозные значения существенно превышают прогноз Минэкономразвития России, что свидетельствует о том, что Минэкономразвития при прогнозировании этого показателя использовал не взаимосвязи между «первичными» и «вторичными» показателями, а некоторые экспертные рассуждения.

Выводы

Научными результатами, представленным в настоящей статье, являются рекуррентные методы вычисления:

1) проекционной матрицы P , порождающей вектор значений линейной модели без оценки её параметров;

2) вектора параметров линейной модели без использования операции обращения матрицы объясняющих переменных.

Прикладное значение метода вычисления проекционной матрицы состоит в том, что на его основе создан алгоритм предназначенный для использования в процедурах поиска наилучшего состава регрессоров MLR-модели в заданном множестве потенциальных аргументов x_1, x_2, \dots, x_p , ориентированной на решение задачи прогнозирования. На первом этапе этих процедур отыскиваются регрессии лучшие по RSS – остаточной сумме квадратов с одним, двумя, тремя и так далее p регрес-

сорами. При этом не требуется параметрическая идентификация модели. Затем среди найденных моделей находится наиболее устойчивое описание, прогнозирующие способности которой наиболее существенны.

Использование предложенного алгоритма на первом этапе процедуры поиска наилучшей модели в некоторых случаях позволяет сократить время и повысить точность вычислений. Указанные эффекты достигаются за счёт увеличения используемых объёмов оперативной памяти.

Предложенный в настоящей статье алгоритм вычисления параметров линейной модели без использования операции обращения матрицы объясняющих переменных представляет, видимо, чисто теоретический интерес, однако, можно утверждать, что изложенные материалы в своей совокупности позволяют исследователям глубже проникнуть в тонкости процедур оценки параметров линейной модели и расширяют арсенал средств аналитиков-практиков. ■

Литература

1. Cochran W.G. The omission or addition an independent variable in multiple linear regression // J. R. Stat. Soc. Suppl., № 5, pp. 171–176.
2. Quenouille M.H. An application of least squares to family diet surveys // Econometrica, № 18, pp. 27–44.
3. Seber G.A.F. Linear regression analysis. Wiley: New York, London, Sydney and Toronto. 1977.
4. Себер Дж. Линейный регрессионный анализ: Пер с англ. М.: Мир, 1980. 456 с.
5. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: в 2-х кн. Кн.1: Пер. с англ. М.: Финансы и статистика, 1986. 366 с.