МЕТОДЫ АНАЛИЗА ДАННЫХ

И.И. Елисеева, С.В. Курышева
(*Санкт-Петербург*)

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ В АНАЛИЗЕ ДАННЫХ

В статье рассматриваются познавательные возможности фиктивных переменных в процедурах анализа взаимосвязи социально-экономических переменных. Основное внимание уделяется трем процедурам: регрессионному анализу, анализу временных рядов и таблиц сопряженности. Роль фиктивных переменных иллюстрируется на примере решения различного класса содержательных задач.

Ключевые слова: номинальные переменные, фиктивная переменная, взаимосвязь переменных, регрессионный анализ, таблица сопряженности, временной ряд.

Постановка исследовательской задачи

В социально-экономических исследованиях задачи выявления характера связи между изучаемыми переменными приходится решать в ситуации, когда они имеют различный уровень измерения (интервальный, порядковый, номинальный). Как правило, *номинальные переменные* используются для того, чтобы отразить структуру исходных для анализа данных и ее влияние на характер взаимосвязи между переменными с более высоким уровнем из-

Ирина Ильинична Елисеева – член-корреспондент РАН, директор Социологического института РАН. E-mail: irinaeliseeva@mail.ru.
Светлана Владимировна Курышева – доктор экономических наук, профессор Санкт-Петербургского государственного университета экономики и финансов. E-mail: stat@finec.ru.

мерения. Так, например, можно ожидать, что зависимость потребления от дохода будет проявляться по-разному у мужчин и у женщин, варьировать по регионам с различной степенью развитости социальной инфраструктуры. На характер этой зависимости могут оказывать влияние и исторические события, и экономическое положение в стране (вероятно, что в период экономического кризиса эффект влияния доходов на потребление товаров длительного пользования будет ниже, чем в стабильной экономике).

В процессе анализа взаимосвязей переменных возникает необходимость во введении такого рода *номинальных переменных* непосредственно в математическую модель. Для этого используются процедуры так называемой оцифровки, когда переменным присваиваются цифровые метки по определенным правилам. Тем самым появляются *фиктивные переменные*. Один из способов их введения следующий: если исходная переменная (мы ее в статье будем называть номинальной) является дихотомической, то формируется одна фиктивная, принимающая два значения: $z = 1$ – наличие какого-либо признака и $z = 0$ – его отсутствие. В случае трех градации (например, для трех уровней образования: высшее, среднее специальное и общее среднее) в модель вводятся две фиктивные переменные z_1 и z_2. При этом $z_1 = 1$ для высшего образования и $z_1 = 0$ для иного уровня образования; $z_2 = 1$ для среднего специального образования и $z_2 = 0$ для иного уровня образования. Если номинальная переменная имеет κ градаций, то в модель вводятся $\kappa - 1$ фиктивных переменных.

Введение в модель фиктивных переменных может преследовать разные цели:

– оценка различий, возникающих в моделируемом показателе за счет особенностей отдельных единиц совокупности при неизменном влиянии других переменных – факторов;

– оценка влияния структурных различий единиц совокупности на характер зависимости объясняемой переменной от объясняющих переменных (факторов), измеренных по интервальной или порядковой шкале;

– оценка влияния номинальной переменной на моделируемый показатель.

Соответственно этим целями фиктивные переменные используются в регрессионном анализе, анализе таблиц сопряженности и динамических рядов. В этой связи последовательно рассмотрим познавательные возможности фиктивных переменных в рамках этих видов анализа.

Регрессионные модели с фиктивными переменными

В случае простого (с одной независимой переменной) линейного регрессионного анализа фиктивные переменные вводятся следующим образом:

$$y = a + bx + cz + e \quad (1)$$

$$y = a + bx + cz + d(xz) + e, \quad (2)$$

где y – зависимая переменная с интервальным уровнем измерения; x – независимая переменная (объясняющий фактор); z – фиктивная переменная, соответствующая двум градациям качественного признака.

При большом количестве объясняющих переменных модель (1) принимает вид:

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + c_1z_1 + c_2z_2 + \ldots + c_mz_m + e. \quad (3)$$

В ней используется $к$ объясняющих переменных (x) и m фиктивных переменных (z), которые могут порождаться несколькими номинальными переменными. Например, в модель необходимо включить три номинальные переменные:

– профессиональные группы (пять категорий);

– образование (три категории);

– пол (две категории).

В этом случае общее число вводимых в модель фиктивных переменных равно семи: $m = 4 + 2 + 1 = 7$. Фиктивные переменные, соответствующие различным номинальным переменным, целесообразно обозначать по-разному: например, z – профессия,

s – образование, v – пол. Тогда в модель будут введены переменные: z_1, z_2, z_3, z_4, s_1, s_2 и v – всего семь фиктивных переменных.

Уменьшение на единицу числа вводимых в модель фиктивных переменных по сравнению с числом градаций номинальной переменной связано со стремлением не попасть в «ловушку» при оценивании параметров регрессии. Если число таких переменных будет соответствовать числу градаций номинальной переменной, то матрица исходных данных станет вырожденной и оценка параметров модели регрессии окажется невозможной [1, с. 169].

Рассмотрим содержательную сторону модели (1). Предположим, что она строится для изучения зависимости потребления кофе от цены. При этом желательно оценить различия в потреблении кофе мужчинами и женщинами. При большом объеме исходных для анализа данных можно построить две модели:

$$y_1 = a + bx + e_1 \text{ – для мужчин;}$$
$$y_2 = A + BX + e_2 \text{ – для женщин,}$$

где y – потребление кофе, x – цена.

Предположим, что параметры этих моделей оказались следующими:

$$y_1 = 120 - 1{,}8x + e_1,$$
$$y_2 = 125 - 1{,}7X + e_2.$$

Это означает, что с ростом цены кофе на одну денежную единицу потребление кофе как мужчинами, так и женщинами снижается в среднем на 1,7-1,8 г. Далее по критерию Г. Чоу проверяется гипотеза: существенно ли различаются коэффициенты регрессии в рассматриваемых уравнениях: $b = -1{,}8$ и $B = -1{,}7$. Если нулевая гипотеза верна H_0 : $b = B = \beta$, т. е. коэффициенты регрессии можно считать одинаковыми, обе выборки можно объединить и построить общую модель регрессии вида (1), в которую введена фиктивная переменная z, принимающая значения, например, $z = 1$ для мужчин и $z = 0$ для женщин. Предположим, что для объединенной совокупности уравнение регрессии составило:

$$y = 126 - 1{,}75X - 7z + e.$$

Значения параметров этого уравнения показывают, что независимо от пола с ростом цены на одну денежную единицу потребление кофе снижается в среднем на 1,75 единиц. Различие в потреблении кофе мужчинами и женщинами отражает коэффициент при $z = -7$. Поскольку z принимает значение 1 для мужчин, то, следовательно, мужчины потребляют кофе на 7 единиц меньше, чем женщины. Подставив в данное уравнение значения z, получим модели потребления кофе для лиц разного пола:

$y = (a - c) - bx + e$ – для мужчин, $y = 119 - 1{,}75x + e$;

$y = a - bx + e$ – для женщин, $y = 126 - 1{,}75x + e$.

На графике эти два уравнения будут представлять собой параллельные линии, что и обусловило название модели (1) – модель регрессии с фиктивными переменными сдвига (см. рис. 1).

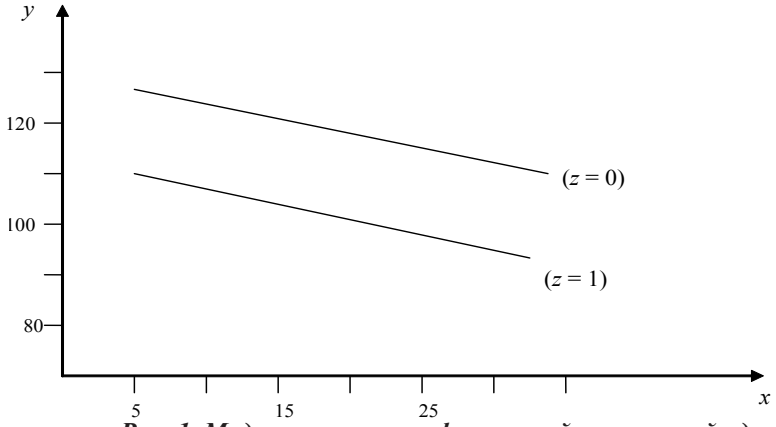

Рис. 1. Модель регрессии с фиктивной переменной сдвига

График показывает, что влияние фактора x на моделируемый показатель y одинаково при разных значениях переменной z. Различен лишь уровень потребления: он выше при $z = 0$.

Если есть основание предположить, что влияние фактора x неодинаково при разных значениях переменной z, то строится модель (2): $y = a + bx + cz + d(xz) + e$.

Соответственно, при $z = 1$ модель примет вид:

$y = (a + c) + (b + d)x + e$, а при $z = 0$: $y = a + bx + e$.

В модель (2) включена еще переменная (xz), коэффициент при ней d характеризует взаимодействие факторов x и z, показывая на сколько единиц изменяется влияние фактора x на y при разных значениях переменной z. Если в модели (1) сила влияния фактора x была одинакова при $z = 1$ и $z = 0$, то в модели (2) влияние фактора x на y различно: при $z = 1$ оно равно $b + d$, а при $z = 0$ оно равно b. Чем больше параметр d, тем значительней различается мера воздействия фактора x на y при разных значениях z. Если при этом окажется, что $d < 0$ и $|d| > |b|$, то влияние фактора x на результат y будет противоположным при $z = 1$ и $z = 0$. Предположим, что модель зависимости спроса на товар А от дохода на одного члена семьи характеризуется уравнением:

$$y = 111 + 2x + 10z - 6(xz) + e,$$

где y – спрос на товар (количество единиц); x – доход на 1 члена семьи (денежных единиц); $z = 1$, если глава семьи работает в бюджетной сфере (образование, здравоохранение); $z = 0$ – если глава семьи работает в финансовой организации (банк, консалтинговая группа и т. д.).

Пусть $z = 1$, тогда модель принимает вид: $y = 121 - 4x + e$, при $z = 0$ имеем: $y = 111 + 2x + e$.

Таким образом, при $z = 1$ связь обратная, а при $z = 0$ – прямая, что на графике соответствует двум перекрещивающимся линиям. Меняется не только направление связи, но и угол наклона линии регрессии (см. рис. 2).

Модель (2) является общим видом модели регрессии с фиктивными переменными. При $к$ объясняющих количественных переменных и $m + 1$ градаций номинальных переменных модель (2) принимает вид:

$$y = a + \sum_{j=1}^{k} b_j x_j + \sum_{i=1}^{m} c_i z_i + \sum_{ji=1}^{km} d_{ji} (x_j z_i) + e. \qquad (4)$$

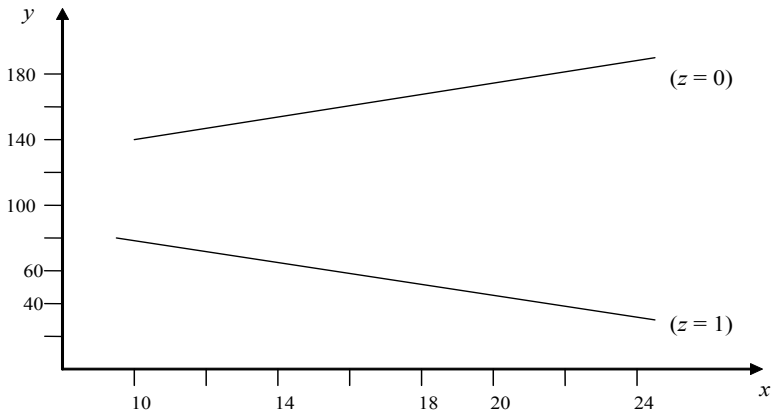

Рис. 2. Модель регрессии с фиктивной переменной, влияющей на направление связи x и y

Построение модели (4) требует достаточно большого объема выборки. Чтобы параметры модели b_j, c_i и d_{ji} были статистически значимыми, необходимо, чтобы на каждый из них приходилось не менее 6-7 наблюдений. Так, при $j = 3$, $m = 2$ модель будет включать 11 переменных, параметры которых могут быть статистически значимыми, если объем выборки составит не менее 70-80 единиц.

Если параметры d_{ji} в модели (4) оказываются статистически незначимыми, то модель (4) превращается в модель (3), в которой параметры $c_1, c_2, \ldots, c_k$ отражают различия в уровнях моделируемого показателя при качественных состояниях отдельных единиц совокупности, представленных в модели фиктивными переменными z.

Если в модели (2) параметр c окажется статистически незначимым, то модель примет вид:

$$y = a + bx + d(xz) + e. \qquad (5)$$

Соответственно, при $z = 1$ модель имеет вид:

$$y = a + (b + d)x + e,$$

а при $z = 0$:

$$y = a + bx + e.$$

Иными словами, модель (5) в величине параметра d фиксирует влияние фиктивной переменной z на меру воздействия фактора x на y, т.е. характеризует изменения в коэффициенте регрессии: $(b + d)$ при $z = 1$ и (b) при $z = 0$. Модель (5) получила название модели *регрессии с фиктивной переменной наклона*. В общем виде при нескольких объясняющих переменных и нескольких фиктивных переменных такая модель регрессии может быть получена из модели (4), в том случае, если параметры сдвига c_i не существенны, т.е. модель примет вид:

$$y = a + \sum_{j=1}^{k} b_j x_j + \sum_{ji}^{k_m} d_{ji} (x_j z_i). \qquad (6)$$

В исследованиях обычно применяют модели регрессии с фиктивными переменными как сдвига, так и наклона, т.е. используют модель (4). Если при этом некоторые из коэффициентов модели b_j, c_i, d_{ij} оказываются статистически незначимыми, то переменные при этих коэффициентах могут быть исключены из модели.

В рассматриваемых моделях фиктивные переменные включались в регрессию наряду с количественными объясняющими переменными (x). Вместе с тем возможно построение модели регрессии только на фиктивных переменных, тогда модель примет вид:

$$y = a + b_1 z_1 + b_2 z_2 + \ldots + b_m z_m + e, \qquad (7)$$

где y – значения моделируемого показателя; $z_1, z_2, \ldots, z_m$ – фиктивные переменные.

В модели (7) под z подразумевается один качественный признак (например, профессия) с $m + 1$ числом категорий (групп), а также несколько нечисловых переменных.

Анализ различий групповых средних

По своему содержанию модель (7) представляет собой аналог аналитической или типологической группировки, в которой па-

раметры при фиктивных переменных фиксируют меру различия среднего значения моделируемого показателя для рассматриваемой группы, для которой $z_j = 1$ относительно другой группы, для которой любое из $z_j = 0$, т.е. $b_j = \bar{y}_j - \bar{y}_0$, где b_j – коэффициент при переменной z_j; $\bar{y}_j$ – среднее значение y по группе j, для которой $z_j = 1$; $\bar{y}_0$ – среднее значение y по базовой группе, с которой ведется сравнение и для которой все $z = 0$, что соответствует значению параметра a.

В этом нетрудно убедиться, применяя для оценки параметров модели (7) метод наименьших квадратов. Пусть совокупность опрошенных делится на три группы: I – удовлетворен профессией; II – не полностью удовлетворен профессией; III – не удовлетворен профессией. По каждой группе опрошенных определен средний доход. Модель регрессии в этом случае примет вид: $y = a + b_1 z_1 + b_2 z_2 + e$, где y – доход (в тыс. руб.); $z_1 = 1$ для лиц первой группы и $z_1 = 0$ для лиц группы II и III; $z_2 = 1$ для лиц второй группы и $z_2 = 0$ для лиц групп I и III.

Следовательно, параметры b_1 и b_2 будут показывать различие средних доходов в группах I и II по сравнению с группой III.

Если для оценки параметров модели $y = a + b_1 z_1 + b_2 z_2 + e$ применить метод наименьших квадратов, то получим систему нормальных уравнений:

$$\begin{cases} \sum y = na + b_1 \sum z_1 + b_2 \sum z_2 \\ \sum y z_1 = a \sum z_1 + b_1 \sum z_1^2 + b_2 \sum z_1 z_2 \\ \sum y z_2 = a \sum z_2 + b_1 \sum z_1 z_2 + b_2 \sum z_2^2 \end{cases}$$

Здесь $\sum z_1 z_2 = 0$; $\sum z_1 = \sum z_1^2 = n_1$ (число наблюдений в группе I); $\sum z_2 = \sum z_2^2 = n_2$ (число наблюдений в группе II); $\sum y z_1 = \sum y_I$ (сумма доходов по группе I); $\sum y z_2 = \sum y_{II}$ (сумма доходов по группе II).

Соответственно, рассматриваемую систему можно записать как:

$$\begin{cases} \sum y = na + b_1 n_1 + b_2 n_2 \\ \sum y_I = n_1 a + n_1 b_1 \\ \sum y_{II} = n_2 a + n_2 b_2 \end{cases}$$

Вычтем из первого уравнения второе и третье, получим:

$$\sum y_{III} = n_3 a,$$

где $\sum y_{III}$ – сумма доходов по группе III, n_3 – число наблюдений в группе III.

Таким образом, $a = \overline{y}_3$, т.е. это средний доход в группе III. Подставим выражение параметра a во второе и третье уравнения:

$\sum y_I = n_1 \overline{y}_3 + n_1 b_1$, отсюда $b_1 = \overline{y}_1 - \overline{y}_3$

$\sum y_{II} = n_2 \overline{y}_3 + n_2 b_2$, отсюда $b_2 = \overline{y}_2 - \overline{y}_3$.

Следовательно, если получено уравнение регрессии $y = 12 + 8z_1 + 3z_2 + e$, то значения параметров показывают, что средний доход в группе I выше на 8 тыс. руб., чем в группе III, а в группе II выше на 3 тыс. руб., чем в группе III, где средний доход составил 12 тыс. руб.

В отличие от других способов группировки модель позволяет получить оценку значимости уравнения в целом с помощью F-критерия Фишера и оценки значимости параметров b_j с помощью t-критерия Стьюдента, т.е. определить существенны ли различия групповых средних. Качество модели оценивается с помощью коэффициента детерминации R^2.

Фиктивные переменные и временные ряды

Анализ временных рядов является еще одним полем использования фиктивных переменных. В качестве иллюстрации рассмотрим задачу построения *аддитивных моделей сезонности*. В такой модели каждое значение признака в период времени t (y_t) представляется в виде суммы трех слагаемых: тренда, сезонности

и случайной компоненты. При этом допустимо, что тренд отсутствует. В этом случае возможны два типа моделей: при отсутствии тенденции; при наличии тенденции во временном ряде.

В любом случае при изучении сезонности рассматриваются данные за ряд лет (не менее трех) по кварталам или месяцам. При отсутствии тенденции в ряде динамики аддитивная модель сезонности при квартальных данных имеет вид:

$$y_t = a + b_1 z_1 + b_2 z_2 + b_3 z_3 + e_t, \qquad (8)$$

где y_t – уровень временного ряда в период времени t (например, объем продажи товара за соответствующий квартал и год); z_1, z_2, z_3 – фиктивные переменные соответственно для кварталов I, II, III, принимающие значение 1 для рассматриваемого квартала и 0 – для остальных кварталов.

В модели (8) все параметры отражают сравнение с кварталом IV, для которого $z_1 = z_2 = z_3 = 0$. Модель (8) аналогична модели (7) с тем лишь отличием, что построена по временному ряду. Интерпретация параметров следующая: a – средний за ряд лет уровень для квартала IV; b_j – разница между средним уровнем j-го квартала и средним уровнем IV квартала.

По модели (8) могут быть найдены средние значения для каждого квартала: $\bar{y}_j = b_j + a$. По ним можно оценить абсолютную величину сезонных колебаний (S_j): $S_j = \bar{y}_j - \bar{y}$, согласно разложению уровня временного ряда на компоненты аддитивной модели:

$$y_t = \bar{y} + S + e, \qquad (9)$$

где y_t – уровень динамического ряда во время t; $\bar{y}$ – средний уровень динамического ряда (например, в среднем за квартал); S – сезонная составляющая, измеренная в тех же единицах, что и уровень ряда; e – случайная компонента, измеренная в тех же единицах, что и уровень ряда.

Предположим, что на основе модели (8) получено уравнение динамики объема продаж:

$$y_t = 30 - 3z_1 + 96z_2 + 144z_3 + e_t.$$

Согласно этому уравнению, средний объем продаж в I квартале был на 3 единицы ниже, а во II и III кварталах на 96 и 144 единиц,

соответственно, выше, чем в IV квартале, когда он составил 30 единиц. Иначе говоря, средний объем продаж составил: в I квартале – 27 единиц, во II – 126, в III – 174, в IV – 30. Среднеквартальный уровень объема продаж равен средней арифметической из средних квартальных уровней, т.е. $\bar{y}$ =89,25 единиц. Тогда показатели сезонных колебаний составят:

I квартал: 27 – 89,25 = –62,25 единиц;
II квартал: 126 – 89,25 = 36,75 единиц;
III квартал: 174 – 89,25 = 84,75 единиц;
IV квартал: 30 – 89,25 = –59,25 единиц.

Если параметры модели по *t*-критерию Стьюдента оказались статистически значимы, то ее можно использовать для прогнозирования подстановкой в нее значений фиктивных переменных. Поскольку рассматривается модель *без тенденции*, то прогнозные значения будут представлять собой средние величины для каждого квартала. Если в годовых уровнях наблюдается тенденция, то используется модель, учитывающая сезонность с помощью фиктивных переменных, имеющая вид (при квартальных данных):

$$y_t = a + bt + c_1 z_1 + c_2 z_2 + c_3 z_3 + e_t, \qquad (10)$$

где t – фактор времени, принимающий значение 1, 2, … l кварталов и учитывающий тенденцию.

В модели (10), как и в модели (8), сезонность представлена фиктивными переменными z_j, но по ней можно получить модель тенденции для каждого квартала:

$y_t = (a+c_1) + bt_t + e$ (I квартал);
$y_t = (a+c_2) + bt + e$ (II квартал);
$y_t = (a+c_3) + bt + e$ (III квартал);
$y_t = a + bt + e$ (IV квартал).

Параметры при фиктивных переменных (c_1, c_2 и c_3) характеризуют изменение уровня соответствующего квартала по сравнению с четвертым кварталом. Параметр b отражает влияние тенденции при элиминировании сезонности. Параметр a фиксирует уровень IV квартала года, предшествующего рассматриваемому периоду времени, т.е. при $t = 0$.

Предположим, что модель (10) использована для характеристики динамики численности безработных (y_t – тыс. чел.) в регионе и имеет следующие значения параметров [2, с. 178]:

$$y_t = 12{,}42 - 0{,}34t - 2{,}03z_1 - 3{,}69z_2 - 5{,}01z_3 + e_t.$$

Модель указывает на тенденцию снижения численности безработных при элиминировании сезонности ежеквартально в среднем на 0,34 тыс. чел. Независимо от действия тенденции уровни ряда в I, II и III кварталах были в среднем ниже, чем в IV квартале ($c_1, c_2, c_3 < 0$). Если в эту модель подставить значения t (от 1 до 12, если данные за три года) и элиминировать фактор сезонности, принимая z_1, z_2, $z_3 = 0$, т.е. на уровне IV квартала, то получим условные значения численности безработных без учета сезонности. Для I квартала первого года ($t = 1$) она составит 12,08 тыс. чел., второго года ($t = 5$) – 10,72 тыс. чел., третьего года ($t = 9$) – 9,36 тыс. чел.

Модель позволяет рассчитать показатели сезонности. Ввиду того, что сумма сезонных компонент $\sum_{j=1}^{4} S_j = 0$, где S_j – показатель сезонности для j-го квартала, $c_j = S_j - S_4$ (характеристика того, насколько показатель сезонности j-го квартала отличается от IV квартала), тогда $S_4 = -\frac{1}{4}(c_1 + c_2 + c_3)$ – сезонная компонента для IV квартала. В нашем примере ее значение равно $S_4 = -\frac{1}{4}(-2{,}03 - 3{,}69 - 5{,}01) = 2{,}682$ тыс. чел. Соответственно, сезонные компоненты для I, II и III кварталов равны:

$S_1 = c_1 + S_4 = -2{,}03 + 2{,}68 = 0{,}65$;
$S_2 = c_2 + S_4 = -3{,}69 + 2{,}68 = -1{,}01$;
$S_3 = c_3 + S_4 = -5{,}01 + 2{,}68 = -2{,}33$;
$S_1 + S_2 + S_3 + S_4 = 0$.

Модель (10) позволяет получить прогнозное значение численности безработных в соответствующем квартале четвертого года. Так, в I квартале четвертого года получим ($t = 13$): $y = 12{,}42 - 0{,}34 \times 13 - 2{,}03 = 5{,}97$. Аналогично рассчитываются значения численности безработных в остальных кварталах.

Мы рассмотрели аддитивную модель сезонной компоненты. При изучении сезонных колебаний довольно часто используются мультипликативные модели, в которых уровень временного ряда представлен как произведение компонент:

$$y_t = \hat{y}_t \cdot K_S \cdot E_t, \qquad (11)$$

где y_t – фактический уровень ряда для времени t; $\hat{y}_t$ – теоретический уровень (тренд) для времени t; K_s – коэффициент сезонности; E_t – коэффициент случайной компоненты.

В отличие от аддитивной модели в модели (11) сезонная составляющая представлена в виде относительной величины – коэффициента сезонности, что приводит при наличии тенденции к меняющейся амплитуде сезонных колебаний: увеличивающейся при $K_s > 1$ и уменьшающейся при $K_s < 1$. Можно ввести в мультипликативную модель и фиктивные переменные. Тогда она принимает вид:

$$y_t = ab^t c_1^{z_1} c_2^{z_2} c_3^{z_3} E_t, \qquad (12)$$

где y_t – фактический уровень; t – фактор времени, учитывающий влияние тенденции (выражается рядом натуральных чисел); z_1, z_2, z_3 – фиктивные переменные (аналогично модели (10)).

Прологарифмировав, получим линейную модель:

$$\ln y_t = \ln a + t \ln b + z_1 \ln c_1 + z_2 \ln c_2 + z_3 \ln c_3 + \ln E_t. \qquad (13)$$

Параметры этой модели определяются методом наименьших квадратов.

Предположим, что динамика потребления товара k в регионе за последние три года в поквартальном разрезе характеризуется уравнением:

$$\ln y_{tj} = 2{,}1469 + 0{,}1267t + 0{,}4714z_1 + 0{,}3486z_2 + 0{,}2652z_3 + \ln E_t,$$

где y_{tj} – потребление товара в квартале j; $t = 1, 2, 3, \ldots, 12$; $z_1 = 1$ для I квартала и 0 – для остальных; $z_2 = 1$ для II квартала и 0 – для остальных; $z_3 = 1$ для III квартала и 0 – для остальных.

Далее путем потенцирования переходим к модели (12), т.е. получаем:

$$a = \text{I}^{2,1469} = 8,558;\ b = \text{I}^{0,1267} = 1,135;\ c_1 = \text{I}^{-0,4714} = 0,624;$$
$$c_2 = \text{I}^{-0,3486} = 0,7057;\ c_3 = \text{I}^{-0,2652} = 0,767.$$

Соответственно, модель (12) запишется как:

$$y_t = 8,558 \cdot 1,135^t \cdot 0,624^{z_1} \cdot 0,7057^{z_2} \cdot 0,767^{z_3} \cdot E_t.$$

Из нее следует, что в динамическом ряду имеется четкая тенденция: ежеквартально независимо от влияния сезонности потребление товара возрастает в среднем на 13,5%. Параметры при фиктивных переменных (c_1 = 0,624; c_2 = 0,7057; c_3 = 0,767) показывают соотношение объема потребления товара в соответствующем квартале к потреблению в IV квартале, принятому за базу сравнения. Поскольку потребление товара в IV квартале было выше, чем в остальных кварталах, то значения параметров $\{c_1, c_2, c_3\} < 1$. Иными словами, в модели (12) параметр b представляет собой средний коэффициент роста уровня ряда динамики независимо от воздействия сезонности, а параметр c_j – влияние сезонности j-го квартала по отношению к IV кварталу независимо от тенденции временного ряда.

Подставляя в модель (12) соответствующие значения z_1, z_2 и z_3, получаем модель динамики объема потребления товара для каждого квартала:

в I квартале: $y_t = (ac_1)\, b^t E_t$;
во II квартале: $y_t = (ac_2)\, b^t E_t$;
в III квартале: $y_t = (ac_3)\, b^t E_t$;
в IV квартале: $y_t = a\, b^t E_t$.

В рассматриваемом примере получены решения этих моделей:
для I квартала: $y_t = (8,558 \cdot 0,624) \cdot 1,135^t E_t$;
для II квартала: $y_t = (8,558 \cdot 0,7057) \cdot 1,135^t E_t$;
для III квартала: $y_t = (8,558 \cdot 0,767) \cdot 1,135^t E_t$;
для IV квартала: $y_t = 8,558 \cdot 1,135^t$.

Все параметры модели (13) оказались статистически значимы по t-критерию Стьюдента, а величина F-критерия составила 94,3,

что характеризует значимость модели в целом: ошибка составила всего 3,6 E – 0,6. Коэффициент детерминации равен 0,982. Это означает, что модель описывает 98,2% вариации исходных данных.

Параметры модели (12) позволяют оценить для каждого квартала коэффициенты сезонности. Известно, что при квартальных данных сумма коэффициентов сезонности должна быть равна 4, $\sum c_j = 4$. Полагая, что параметр c_j отражает влияние сезонности j-го квартала по отношению к IV кварталу, получаем для нашего примера:

$$c_1 + c_2 + c_3 + 1 = 3{,}0967.$$

Для оценки коэффициентов сезонности найдем поправочный коэффициент: $K_{\text{поправки}} = {}^{4}/_{3{,}0967} = 1{,}2916976$.

Умножая его на значения параметров c_1, c_2 и c_3, получаем поквартальные коэффициенты сезонности (K_{s_j}) :

для I квартала: 0,806;

для II квартала: 0,911;

для III квартала: 0,991;

для IV квартала: 1,292.

Сумма скорректированных коэффициентов сезонности равна 4.

Коэффициенты сезонности имеют аналитическое значение. Для прогнозирования объема потребления достаточно пользоваться параметрами модели (12).

Фиктивные переменные и анализ таблиц сопряженности

Таблицы сопряженности являются основой для изучения взаимосвязи переменных с низким уровнем измерения (номинальный, порядковый). Существуют различные подходы к анализу таких таблиц [3]. Один из них включает использование фиктивных переменных. Поясним его суть на примере таблицы сопряженности, отражающей распределение рабочих высокой и средней квалификации в зависимости от видов труда с точки зрения его механизации (табл. 1).

Таблица 1

РАСПРЕДЕЛЕНИЕ РАБОЧИХ ПО ВИДАМ ТРУДА,
% к числу наблюдений

Квалификация	Вид труда		Итого
	Механизированный	Ручной	
Высокая	17,1	25,1	42,2
Средняя	12,1	45,7	57,8
Итого	29,2	70,8	100

В таблице представлено двумерное распределение, которое описывается *моделью распределения*:

$$z = b_0 + b_1x + b_2y + b_3xy, \qquad (14)$$

где z_{ij} – удельный вес рабочих i-й квалификации и j-го вида труда; x_j – вид труда: $x_j = 1$ – механизированный труд; $x_j = 0$ – ручной труд; y_i – квалификация: $y_i = 1$ – высокая; $y_i = 0$ – средняя.

В такой модели x и y являются фиктивными переменными со значениями 1 и 0. Для ее построения таблица сопряженности должна быть представлена в виде матрицы данных (табл. 2).

Таблица 2

ТАБЛИЦА СОПРЯЖЕННОСТИ С ФИКТИВНЫМИ ПЕРЕМЕННЫМИ

x	y	xy	z
1	1	1	17,1
1	0	0	12,1
0	1	0	25,1
0	0	0	45,7

Методом наименьших квадратов получим модель распределения рабочих: $z = 45{,}7 - 33{,}6x - 20{,}6y + 25{,}6xy$.

Подставляя значения фиктивных переменных, получим распределение численности рабочих, соответствующее рассматриваемой таблице сопряженности:

при $x = y = 0$ $z = 45{,}7$;

при $x = 0,\ y = 1 \quad z = 25{,}1;$
при $x = 1,\ y = 0 \quad z = 12{,}1;$
при $x = 1,\ y = 1 \quad z = 17{,}1.$

Иными словами модель (14) выступает аналогом таблицы сопряженности. В ней параметр $b_0 = 45{,}7$ характеризует удельный вес численности рабочих базовой группы, т.е. группы, для которой $x = 0$ и $y = 0$. Параметр $b_1 = -33{,}6$ показывает, на сколько процентных пунктов удельный вес численности рабочих механизированного труда ($x = 1$) и средней квалификации ($y = 0$) оказался ниже, чем удельный вес базовой группы, т.е. $b_1 = 12{,}1 - 45{,}7 = -33{,}6$. Параметр $b_2 = -20{,}6$ также фиксирует, на сколько процентных пунктов удельный вес численности рабочих ручного труда и высокой квалификации ($y = 1, x = 0$) оказался ниже базовой группы, т.е. $b_2 = 25{,}1 - 45{,}7 = -20{,}6$. Параметр $b_3 = 25{,}6$ фиксирует дополнительное изменение удельного веса численности рабочих для сочетания $x = y = 1$. В данной модели $\sum_0^3 b_j = 17{,}1$, т.е. это удельный вес численности рабочих в подгруппе с $x = 1$ и $y = 1$. Эту же величину получим, подставляя в модель значения $x = y = 1$. Тем самым, имея модель (14), можно восстановить всю информацию, имеющуюся в таблице сопряженности.

Рассматриваемая модель распределения не содержит случайной составляющей. Поэтому естественно, что коэффициент детерминации для нее $R^2 = 1$. Его значение не может рассматриваться как показатель тесноты связи между x и y. Для этой цели, как известно, в случае четырехклеточных таблиц используется коэффициент контигенции:

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}},$$

где a, b, c, d – частоты четырехпольной таблицы.

Это выражение коэффициента контингенции получается путем преобразования формулы линейного коэффициента корреляции для фиктивных переменных, когда x и y принимают только значения 1 и 0.

Рассмотренный подход описания таблиц сопряженности типа 2 х 2 может быть распространен и на таблицы большей размерности. Если таблица типа 3 х 3, то для ее описания потребуется 8 фиктивных переменных: по две для признаков x и y, т.е. в модель вводится x_1 и x_2, y_1 и y_2, и, кроме того, четыре переменных для отражения *взаимодействий* x и y. Тогда, модель примет вид:

$$z = a + b_1x_1 + b_2x_2 + c_1y_1 + c_2y_2 + d_1(x_1y_1) + d_2(x_1y_2) + d_3(x_2y_1) + d_4(x_2y_2), \quad (15)$$

где z – числовые данные в клетке таблицы сопряженности; $x_1 = 1$ – для категории I по признаку x; $x_1 = 0$ – для остальных категорий признака x; $y_1 = 1$ – для категории I признака y; $y_1 = 0$ – для остальных категорий; $x_2 = 1$ – для категории II признака x; $x_2 = 0$ – для других категорий признака x; $y_2 = 1$ – для категории II признака y; $y_2 = 0$ – для других категорий признака y; x_1y_1; x_1y_2; x_2y_1; x_2y_2 – соответствующие сочетания категорий x и y.

Для построения модели составляется матрица исходных данных (табл. 3).

Таблица 3

МАТРИЦА ИСХОДНЫХ ДАННЫХ

z								
	x_1	x_2	y_1	y_2	x_1y_1	x_1y_2	x_2y_1	x_2y_2
5,2	1	0	1	0	1	0	0	0
7,3	0	1	1	0	0	0	1	0
12,5	0	0	1	0	0	0	0	0
7,8	1	0	0	1	0	1	0	0
15	0	1	0	1	0	0	0	1
7,2	0	0	0	1	0	0	0	0
12	1	0	0	0	0	0	0	0
20	0	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0

Применяя метод наименьших квадратов, получаем оценки параметров модели – модели клеточных частот:

$$z = 13 + 1x_1 + 7x_2 - 0{,}5y_1 - 5{,}8y_2 - 6{,}3(x_1y_1) + 1{,}6(x_1y_2) - 12{,}2(x_2y_1) + 0{,}8(x_2y_2).$$

По ней восстанавливается и сама таблица сопряженности (табл. 4).

Таблица 4

РАСПРЕДЕЛЕНИЕ, ВОССТАНОВЛЕННОЕ ПО МОДЕЛИ

	y_1	y_2	y_3
x_1	5,2	7,8	12
x_2	7,3	15	20
x_3	12,5	7,2	13

В этом можно убедиться, проанализировав модель (15). В ней группы со значениями x_3 и y_3 взяты за базу сравнения, для них всегда $y_1 = y_2 = 0$ и $x_1 = x_2 = 0$. Параметр a соответствует удельному весу (в %) численности единиц совокупности в клетке таблицы, соответствующей сочетанию x_3, y_3, ибо для этой позиции все переменные в модели равны нулю. Эта величина выступает базой сравнения для удельных весов в других клетках таблицы. Так, например, $b_1 = -1$ показывает, на сколько процентных пунктов удельный вес в клетке x_1 и y_3 меньше, чем в базовой клетке. Подставляя в модель $x_1 = 1$ и остальные переменные, равные 0, получим $z = 12$, что соответствует сочетанию x_1y_3. Аналогично трактуются параметры b_2, c_1, c_2. Они позволяют оценить удельные веса в клетках x_2y_3 (при $b_2 = 7$, $z = 20$), y_1x_3 (при $c_1 = -0{,}5$, $z = 12{,}5$), y_2x_3 (при $c_2 = -5{,}8$, $z = 7{,}2$).

Параметры d_1, d_2, d_3, d_4 фиксируют дополнительное изменение удельного веса для конкретных сочетаний x и y в клетках таблицы. Так, $d_1 = -6{,}3$ означает, что эта величина должна быть прибавлена при расчете удельного веса для клетки таблицы x_1y_1. Подставляя в модель значения 0 и 1 для клетки x_1y_1, получим $z = a + b_1 + c_1 + d_1 = 13 - 1 - 0{,}5 - 6{,}3 = 5{,}2$. Соответственно в клетке x_1y_2 появляется число $z = a + b_1 + c_2 + d_2 = 13 - 1 - 5{,}8 + 1{,}6 = 7{,}8$. Для сочетания

x_2y_1 получим $z = a + b_2 + c_1 + d_3 = 13 + 7 – 0,5 – 12,2 = 7,3$. Для сочетания x_2y_2 удельный вес в таблице сопряженности составит: $z = a + b_2 + c_2 + d_4 = 13 + 7 – 5,8 + 0,8 = 15$.

Для модели (15) коэффициент детерминации равен 1, что соответствует полному распределению единиц совокупности. Используя матрицу исходных данных, можно проанализировать матрицу коэффициентов корреляции и определить, какое из сочетаний признаков *x* и *y* в большей мере сказывается на фактическом распределении единиц совокупности.

Таким образом, приведенные примеры показывают, что фиктивные переменные имеют широкий спектр применения в социально-экономических исследованиях. С их помощью можно изучать влияние природных, технологических, социальных, психологических факторов на количественные изменения социально-экономических показателей. Следует подчеркнуть, что фиктивные переменные могут выступать и в качестве зависимых в отличие от независимых (объясняющих) в рассмотренных нами случаях. Тогда строится линейная вероятностная модель или *logit*- и *ptobit*-модели (см. подробнее в [4, с. 299, 330–331]).

ЛИТЕРАТУРА

1. Эконометрика: Учебник / Под ред И.И. Елисеевой. 2-е изд. М.: Финансы и статистика, 2005.

2. Эконометрика: Учебник / Под ред. И.И. Елисеевой. М.: Проспект, 2008.

3. *Аптон Г.* Анализ таблиц сопряженности / Пер. с англ. М.: Финансы и статистика, 1982.

4. *Вербик М.* Путеводитель по современной эконометрике / Науч. ред. и предисл. докт. физ.-мат. наук, проф. С.А. Айвазяна. М.: Научная книга, 2008. (Библиотека Солев.)