УДК 519.233.4+519.237.4

С.И. Жилин

Решение задач дисперсионного и ковариационного анализа методом центра неопределенности

S.I. Zhilin

Solving Problems of Dispersion and Covariation Analysis Using Uncertainty Center Method

Предложен способ решения задач дисперсионного и ковариационного анализа в рамках интервального нестатистического подхода к описанию неопределенности в данных. Способ заключается в сведении задач построения и анализа эмпирических зависимостей при необходимости учета влияния качественных факторов к задаче регрессионного анализа и ее последующем решении методом центра неопределенности.

Ключевые слова: нестатистический подход, интервальное оценивание, дисперсионный анализ, ковариационный анализ, метод центра неопределенности.

The paper describes a technique of ANOVA-type problem solving on the base of interval non-statistical approach to data uncertainty handling. The essence of the technique is to reduce a problem of building and analyzing empirical dependencies under the influence of categorical factors to the problem of regression analysis which could be solved using uncertainty center method.

Key words: non-statistical approach, interval estimates, analysis of variance, analysis of covariance, uncertainty center method.

Введение. Интервальный (нестатистический) подход к обработке и анализу экспериментальной информации основывается на описании неопределенностей в данных ограниченными множествами, чаще всего задаваемыми интервалами или их декартовыми произведениями – брусами. При этом на множествах неопределенности не вводится никаких дополнительных мер (вероятностных, нечетких и пр.). Такой взгляд на обрабатываемые данные хорошо согласуется с запросами практиков, зачастую не владеющих информацией о вероятностной структуре этих данных, особенно в случае коротких выборок.

Идейно восходя к пионерской работе Л.В. Канторовича [1] и часто совпадая содержательно, но различаясь терминологически, приемы построения и анализа эмпирических зависимостей на базе интервального подхода развиваются различными группами отечественных и зарубежных исследователей [2–9]. Выработанная техника оценивания параметров и построения прогнозов зависимостей позволяет существенно обогатить сведения, получаемые аналитиком, о восстанавливаемой зависимости и ее свойствах более традиционными статистическими методами. При этом большинство известных результатов в области интервального (нестатистического) подхода касаются постановки задачи, известной как задача регрессионного анализа, которая состоит в поиске и анализе приемлемой модели зависимости между количественными экзогенными и количественной же эндогенной переменными. Однако на практике при построении зависимости часто приходится сталкиваться с необходимостью учета некоторых качественных факторов. Ситуацию, когда экзогенные переменные представлены исключительно качественными факторами, обычно именуют задачей дисперсионного анализа [10, 11]. Задачу же изучения зависимости, в которой наряду с качественными имеются и количественные экзогенные факторы, принято называть задачей ковариационного анализа [10, 12]. В настоящей работе предложен способ решения задач этих двух типов в рамках интервального (нестатистического) подхода. В первом разделе работы изложены основные идеи интервального подхода, при этом используется терминология метода центра неопределенности [4, 5]. Во втором разделе показано, каким образом задачи дисперсионного и ковариационного анализа могут быть решены с помощью метода центра неопределенности. Наконец, в третьем разделе приведен простой численный пример.

1. Метод центра неопределенности. Основу метода центра неопределенности составляет техника исследования множества допустимых значений параметров зависимости, конструируемой по таблице наблюдений за экзогенными и эндогенной переменными. При этом полагается, что значения экзогенных переменных известны точно (или с пренебрежимо малыми погрешностями), а суммарная ошибка наблюдения эндогенной переменной ограничена сверху по модулю величиной ε. В частности, в случае построения линейно-параметризованной зависимости вида

$$y = \sum_{i=0}^{n} \beta_i x_i \qquad (1)$$

по таблице экспериментальных данных, полученной в N наблюдениях,

$$T = \left\{ \left(y_j,\ x_{0j}, x_{1j}, ..., x_{nj} \right) \,\middle|\, j = 1, ..., N \right\} \qquad (2)$$

множество допустимых значений параметров зависимости представляет собой полиэдральное, а следовательно, и выпуклое множество

$$B = \left\{ \beta = (\beta_0, ..., \beta_n) \,\middle|\, y_j - \varepsilon_j \le \sum_{i=0}^{n} \beta_i x_{ij} \le y_j + \varepsilon_j \right\}, \quad j = 1, ..., N. \qquad (3)$$

При этом B ограничено тогда и только тогда, когда ранг матрицы наблюдений $X = (x_{ij})_{(n+1)\times N}$ равен $n + 1$. Содержательно неограниченность множества B может интерпретироваться как недостаток эмпирической информации. Пустота множества B говорит о противоречивости собранной информации.

Главным принципом нестатистической обработки наблюдений, определяющим все последующие алгоритмы и получаемые выводы, является отсутствие каких-либо предпочтений для элементов множества B (их равноправие при выборе в качестве оценок параметров).

Ввиду сложности полного описания множества B в ряде случаев ограничиваются некоторыми его аппроксимациями. В частности, в этой роли можно использовать брусы (гиперпараллелепипеды с гранями, параллельными координатным плоскостям), охватывающие множество неопределенности B. Наименьший из таких брусов отыскивается путем решения следующих задач линейного программирования:

$$\underline{\beta}_i = \min_{\beta \in B} \beta_i, \quad \overline{\beta}_i = \max_{\beta \in B} \beta_i, \quad i = 0, ..., n. \qquad (4)$$

Интервалы $[\underline{\beta}_i, \overline{\beta}_i]$, $i = 0, ..., n$, определяющие этот брус, содержат в себе возможные точечные оценки параметров β_i, а их длины могут выступать в качестве меры точности точечных оценок.

В соответствии с главным принципом нестатистической обработки наблюдений точечной оценкой параметров β_i зависимости (1) в равной степени может служить любой из элементов множества B. Известен ряд подходов к выбору представительной точки из множества B, опирающихся на различные соображения [13], но одним из наиболее простых способов построения точечной оценки $\hat{\beta} = \left(\hat{\beta}_0, ..., \hat{\beta}_n \right)$ является выбор в этом качестве срединной точки охватывающего бруса, отыскиваемого при решении задач (4):

$$\hat{\beta}_i = \frac{1}{2}\left(\underline{\beta}_i + \overline{\beta}_i \right), \ \Delta\beta_i = \frac{\overline{\beta}_i - \underline{\beta}_i}{2}, \ i = 0, ..., n. \qquad (5)$$

Помимо задачи точечного и интервального оценивания параметров зависимости в отношении множества B может ставиться и задача интервального оценивания выходной переменной y зависимости (1) в точке x:

$$\underline{y}(x) = \min_{\beta \in B} \beta x, \quad \overline{y}(x) = \max_{\beta \in B} \beta x. \qquad (6)$$

Интервал $[\underline{y}(x),\ \overline{y}(x)]$ содержит возможные значения выходной переменной y в точке x при различном выборе параметров зависимости. В качестве точечной оценки прогнозного значения зависимости (1) в точке x по аналогии с (5) может использоваться полусумма концов интервала:

$$\hat{y}(x) = \frac{1}{2}\left(\underline{y}(x) + \overline{y}(x) \right), \ \Delta y(x) = \frac{1}{2}\left(\overline{y}(x) - \underline{y}(x) \right). \qquad (7)$$

С использованием гарантированных интервальных оценок параметров и прогнозных значений зависимости довольно просто проводится анализ значимости коэффициентов зависимости [5, 6]. В случае пустоты множества допустимых параметров зависимости B возможно выявление выбросов [14] и/или построение совместных подвыборок наблюдений [15].

Таким образом, базовые приемы метода центра неопределенности позволяют исследователю решать тот же круг вопросов, которые находятся в фокусе классического регрессионного анализа.

2. Задачи ковариационного и дисперсионного анализа. Основной прием, позволяющий при построении зависимости ввести в рассмотрение качественные факторы, состоит в использовании фиктивных переменных. В классическом статистическом анализе хорошо известны [10–12, 16] способы применения этого аппарата для сведения задач дисперсионного анализа (все факторы качественные) и ковариационного анализа (часть факторов – количественные, а часть – качественные) к задаче регрессионного анализа. В настоящем разделе будет показано, что использование того же приема при нестатистическом подходе делает возможным решение задач дисперсионного и ковариационного анализов с помощью метода центра неопределенности.

Для учета влияния на значение выходной переменной каждого из качественных факторов x_i, принимающих значения на L_i уровнях $X_i = \{x_{i0}, ..., x_{iL_i-1}\}$, в зависимость вводятся $L_i - 1$ фиктивных переменных, значения которых в совокупности кодируют уровень фактора x_i, соответствующий каждому из наблюдений. Способ выбора фиктивных переменных не единственен. Одним из наиболее простых для реализации и интерпретации является следующий вариант сопоставления уровней фактора и значений совокупности фиктивных переменных.

Один из уровней фактора выбирается в качестве эталонного, например, x_{i0}, а для остальных определяются фиктивные переменные $d_{i1},...,d_{i(L_i-1)}$, принимающие значения 0 или 1. Ситуация, когда все переменные $d_{i1},...,d_{i(L_i-1)}$ равны нулю, соответствует эталонному уровню фактора x_{i0}. Равенство единице переменной d_{ik} при нулевых значениях остальных фиктивных переменных соответствует уровню фактора x_{ik}.

Коэффициент δ_{ik} при каждой из заданных таким способом фиктивных переменных d_{ik} представляет собой оценку так называемого чистого эффекта, т.е. разницы в значении выходной переменной, обусловленной переходом фактора x_i с эталонного уровня x_{i0} на уровень x_{ik} при фиксированных значениях прочих переменных, входящих в зависимость.

После пополнения фиктивными переменными структура зависимости приобретает вид

$$y = \sum_{i=0}^{m-1} \beta_i x_i + \sum_{i=m}^{n} \sum_{k=1}^{L_i-1} \delta_{ik} d_{ik} , \qquad (8)$$

где входные переменные $x_0,...,x_{m-1}$ являются количественными факторами, а качественные факторы $x_m,...,x_n$ представлены группами фиктивных переменных $d_{i1},...,d_{iL_i-1}$, $i = m,...,n$.

При $m = 0$ задача построения и анализа зависимости вида (2) соответствует задаче дисперсионного анализа, а при $m > 0$ – задаче ковариационного анализа. Для оценивания коэффициентов β_i и δ_{ik} используются методы, изложенные в предыдущем разделе.

3. Пример. Данные для примера (табл., рис.) взяты из [12, с. 301] и представляют собой вес (y) в фунтах и возраст в неделях для 13 индеек. Четыре из них выращены в штате Джорджия, четыре – в Виргинии и пять – в Висконсине. Попытаемся связать вес и возраст птицы простой линейной зависимостью и выяснить, какое влияние на зависимость оказывает место ее происхождения. Для учета влияния этого качественного фактора, принимающего значение на трех уровнях, введем две фиктивные переменные d_1 и d_2, определив их значения. Конструируемая зависимость имеет вид:

$$y = \beta_0 + \beta_1 x + \delta_1 d_1 + \delta_2 d_2 + \varepsilon . \qquad (9)$$

Данные об индейках

Номер опыта	Вес, фунтов (y)	Возраст, недель (x)	Место происхождения	d_1	d_2
1	13,3	28	Джорджия	1	0
2	8,9	20	Джорджия	1	0
3	15,1	32	Джорджия	1	0
4	10,4	22	Джорджия	1	0
5	13,1	29	Виргиния	0	1
6	12,4	27	Виргиния	0	1
7	13,2	28	Виргиния	0	1
8	11,8	26	Виргиния	0	1
9	11,5	21	Висконсин	0	0
10	14,2	27	Висконсин	0	0
11	15,4	29	Висконсин	0	0
12	13,1	23	Висконсин	0	0
13	13,8	25	Висконсин	0	0

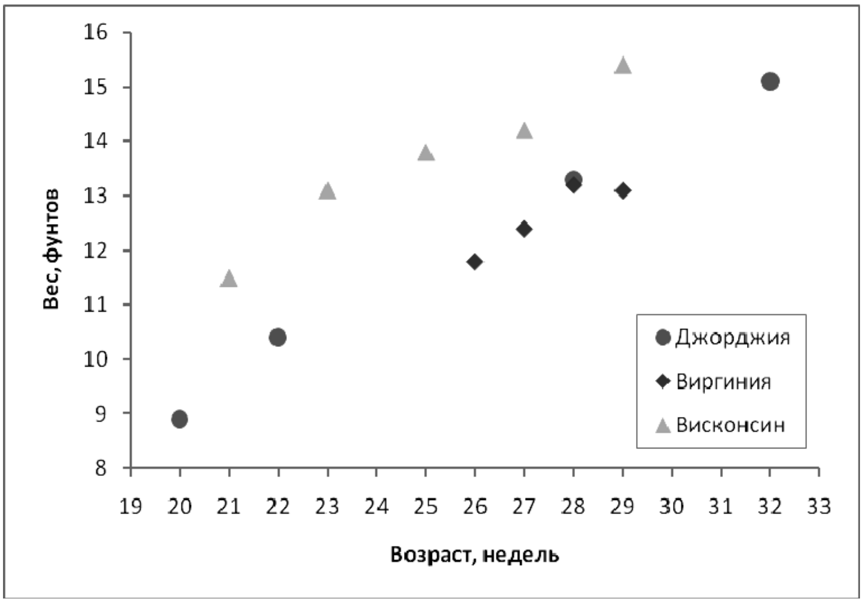

Диаграмма рассеяния для данных об индейках

Отсутствующую в первоисточнике информацию о верхней границе абсолютного значения ошибки ($\bar{\varepsilon}$) измерения выходной переменной восполним, положив ее равной 1 фунту. Множество неопределенности в нашей задаче определяется неравенствами вида

$$y_j - \bar{\varepsilon} \le \beta_0 + \beta_1 x_j + \delta_1 d_{1j} + \delta_2 d_{2j} \le y_j + \bar{\varepsilon},\ j = 1,...,13, \quad (10)$$

где $(y_j, x_j, d_{1j}, d_{2j})$ – данные из таблицы.

Используя те же процедуры метода центра неопределенности, что и при решении задачи регрессионного анализа, получаем интервальные оценки параметров зависимости:

$$\hat{\beta}_0 \in [1,750;\ 5,570],\ \hat{\beta}_1 \in [0,350;\ 0,450],$$

$$\hat{\delta}_1 \in [-3,350; -0,850],\ \hat{\delta}_2 \in [-2,600; -0,450].$$

В качестве точечной оценки примем наиболее просто вычисляемый центр прямоугольника:

$$\hat{\beta}_0 = 3,750,\ \hat{\beta}_1 = 0,400,\ \hat{\delta}_1 = -2,100,\ \hat{\delta}_2 = -1,525.$$

Однозначно отрицательные интервальные оценки коэффициентов δ_1, δ_2 при фиктивных переменных указывают на различия в индейках, первая – из Джорджии и Висконсина, а вторая – из Виргинии и Висконсина соответственно. Подставляя три различных набора значений фиктивных переменных (d_1, d_2) и используя точечные оценки параметров, получим зависимости, описывающие характеристики птиц, выращенных на трех разных территориях:

для Джорджии при $d_1 = 1$, $d_2 = 0$:

$$\hat{y} = 1,650 + 0,400x:$$

для Виргинии при $d_1 = 0$, $d_2 = 1$:

$$\hat{y} = 2,225 + 0,400x;$$

для Висконсина при $d_1 = 0$, $d_2 = 0$:

$$\hat{y} = 3,750 + 0,400x.$$

Полученные результаты не противоречат результатам обработки этих данных классическими методами регрессионного анализа [12].

Заключение. Таким образом, задачи учета влияния качественных факторов при построении и анализе зависимостей по экспериментальным данным, традиционно решаемые статистическими методами дисперсионного и ковариационного анализа, могут быть с успехом решены и в рамках интервального (нестатистического) подхода. Достоинствами интервального подхода являются существенно более простая система условий применимости его методов и естественная для аналитиков-практиков форма представления информации о неопределенности в данных.

Библиографический список

1. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сиб. мат. журнал. – 1962. – Т. 3, №5.

2. Спивак С.И. Информативность эксперимента и проблема неединственности решения обратных задач химической кинетики: автореф. дис. … д-ра физ.-мат. наук. – Черноголовка, 1984.

3. Bounding Approaches to System Identification / Milanese M., Norton J., Walter E., editors. – London, 1996.

4. Белов В.М., Суханов В.А., Унгер Ф.Г. Теоретические и прикладные аспекты метода центра неопределенности. – Новосибирск, 1995.

5. Оскорбин Н.М., Максимов А.В., Жилин С.И. Построение и анализ эмпирических зависимостей методом центра неопределенности // Известия АлтГУ. – 1998. – №1.

6. Вощинин А.П., Бочков А.Ф., Сотиров Г.Р. Метод анализа данных при интервальной нестатистической ошибке // Заводская лаборатория. – 1990. – Т. 56, №7.

7. Померанцев А.Л., Родионова О.Е. Построение многомерной градуировки методом простого интервального оценивания // Жур. аналит. химии. – 2006. – №61.

8. Кумков С.И. Обработка экспериментальных данных ионной проводимости расплавленного электролита методами интервального анализа // Расплавы. – 2010. – №3.

9. Подружко А.А., Подружко А.С. Интервальное представление полиномиальных регрессий. – М., 2003.

10. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. – М., 1976.

11. Шеффе Г. Дисперсионный анализ. – М., 1980.

12. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. – М., 1987.

13. Жилин С.И. Нестатистические модели и методы построения и анализ эмпирических зависимостей: дис. … канд. физ.-мат. наук. – Барнаул, 2004.

14. Zhilin S.I. Simple Method for Outlier Detection in Fitting Experimental Data Under Interval Error // Chemometrics and Intellectual Laboratory Systems. – 2007. – Vol. 88(1).

15. Кумков С.И. Интервальный подход к обработке зашумленных экспериментальных данных с многократными измерениями в условиях неопределенности // Современные проблемы прикладной математики и механики: теория, эксперимент и практика: докл. Междунар. конф., посвящ. 90-летию со дня рождения акад. Н.Н. Яненко. – Новосибирск, 2011.

16. Доугерти К. Введение в эконометрику. – М., 1999.