Математические методы исследования

УДК 519.234

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ ТОЧКИ ПЕРЕСЕЧЕНИЯ РЕГРЕССИОННЫХ ПРЯМЫХ

© **В. С. Муравьева, А. И. Орлов**[1]

Статья поступила 13 декабря 2006 г.

Описан метод доверительного оценивания точки пересечения двух регрессионных линейных зависимостей. В асимптотической постановке изучена непараметрическая вероятностно-статистическая модель (без предположения о нормальности распределения погрешностей). На основе метода линеаризации получены асимптотические дисперсия и доверительный интервал для точки встречи. Метод рассмотрен на примере сравнительного анализа тенденций развития отечественной и зарубежной групп продукции.

Пусть зависимость от времени t некоторого показателя $x_1(t)$ технического уровня или качества продукции предприятия "Альфа" описывается линейной функцией

$$x_1(t) = a_1 t + d_1.$$

Пусть аналогичный показатель у его конкурента ("Бета") также описывается линейной функцией, но с другими коэффициентами:

$$x_2(t) = a_2 t + d_2.$$

Предположим, что предприятие "Альфа" находится в положении догоняющей стороны. Это значит, что в рассматриваемый момент времени t_0 (например, "сегодня") значение показателя его продукции ниже: $x_1(t_0) < x_2(t_0)$, но темп роста у предприятия "Альфа" выше, чем у конкурента: $a_1 > a_2$.

Возникает естественный вопрос: когда предприятие "Альфа" догонит конкурента? Другими словами, в какой момент времени будет выполнено равенство $x_1(t) = x_2(t)$? Решая относительно t уравнение

$$a_1 t + d_1 = a_2 t + d_2,$$

получим, что встреча произойдет в момент

$$t_{\text{в}} = \frac{d_2 - d_1}{a_1 - a_2}.$$

Представляют интерес еще две величины. Во-первых, уровень качества, при котором предприятие "Альфа" сравняется с конкурентом, т.е. общий уровень качества в момент встречи:

$$x = x_1(t_{\text{в}}) = x_2(t_{\text{в}}) = \frac{a_1 d_2 - a_2 d_1}{a_1 - a_2}.$$

Во-вторых, временно́й лаг, т.е. величина отставания предприятия "Альфа" в рассматриваемый момент времени t_0. В какой (более ранний) момент времени t_k конкурент имел тот уровень качества, которого предприятие "Альфа" достигло сейчас? Ответом на этот вопрос будет решение уравнения $x_2(t) = x_1(t_0)$:

$$t_k = \frac{x_1(t_0) - d_2}{a_2}.$$

Следовательно, предприятие "Альфа" отстает на

$$L = t_0 - t_k = \frac{(a_2 - a_1)t_0 + d_2 - d_1}{a_2} = \frac{x_2(t_0) - x_1(t_0)}{a_2}$$

единиц времени (лет).

В реальных ситуациях линейные зависимости неизвестны. Однако известны исходные данные $(t_{i1}; x_{i1})$, $i = 1, 2, ..., n(1)$ — для предприятия "Альфа" и $(t_{j2}; x_{j2})$, $j = 1, 2, ..., n(2)$ — для предприятия-конкурента. При этом значения показателя $x_1(t_{i1}) = x_{i1}$ у предприятия "Альфа" в моменты времени t_{i1} представляются в виде

$$x_1(t_{i1}) = x_{i1} = a_1 t_{i1} + d_1 + e_{i1}, \ i = 1,2, ..., n(1),$$

где коэффициенты a_1 и d_1 неизвестны статистику, а e_{i1} — погрешности измерения (невязки). Будем считать, что e_{i1}, $i = 1, 2, ..., n(1)$ — совокупность незави-

[1] Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия.

симых одинаково распределенных случайных величин с нулевым математическим ожиданием и дисперсией $D(e_{i1}) = \sigma_1^2$, не известной статистику.

Для предприятия-конкурента справедливо аналогичное представление:

$$x_2(t_{j2}) = x_{j2} = a_2 t_{j2} + d_2 + e_{j2},\ j = 1, 2, ..., n(2),$$

где коэффициенты a_2 и d_2 неизвестны статистику, а e_{j2} — погрешности измерения (невязки). Примем, что e_{j2}, $j = 1, 2, ..., n(2)$ — совокупность независимых одинаково распределенных случайных величин с нулевым математическим ожиданием и дисперсией $D(e_{j2}) = \sigma_2^2$, не известной статистику.

Примем, что совокупности случайных величин e_{i1}, $i = 1, 2, ..., n(1)$ и e_{j2}, $j = 1, 2, ..., n(2)$ независимы между собой. В каждой совокупности случайные величины одинаково распределены, но функции распределения, соответствующие разным совокупностям (т.е. предприятиям "Альфа" и "Бета"), могут различаться между собой.

В рассматриваемой вероятностно-статистической модели не предполагается, что эти функции распределения входят в какое-либо параметрическое семейство распределений (в частности, что невязки имеют нормальное распределение). А это значит, что рассматривается непараметрическая постановка. Однако будем считать, что объемы данных $n(1)$ и $n(2)$ достаточно велики, поэтому можно применять центральную предельную теорему и приближать совместное распределение оценок метода наименьших квадратов с помощью многомерного нормального распределения.

Итак, решение задачи о точке встречи получим в рамках непараметрической вероятностно-статистической модели. Частный случай, когда невязки e_{i1}, $i = 1, 2, ..., n(1)$ и e_{j2}, $j = 1, 2, ..., n(2)$ имеют нормальное распределение, приведен в работе [1].

Рассмотрим метод решения задачи о встрече. Вместо неизвестных статистику зависимостей $x_1(t)$ и $x_2(t)$ будем использовать их оценки $x_1^*(t)$ и $x_2^*(t)$, полученные методом наименьших квадратов. Для этого необходимо оценить коэффициенты по правилам, полученным в работе [2, п. 9.2], а затем рассчитать оценки момента встречи $t_в^*$, уровня качества в момент встречи $x^* = x_1^*(t_в^*) = x_2^*(t_в^*)$ и временно́го лага (величины отставания)

$$L^* = \frac{x_2^*(t_0) - x_1^*(t_0)}{a_2^*},$$

используя оценки коэффициентов зависимостей $x_1(t)$ и $x_2(t)$ вместо неизвестных истинных коэффициентов.

Полезным является, как и в работе [2], использование центрирования средними значениями независимой переменной при параметризации зависимостей:

$$x_1(t) = a_1[t - t_{ср}(1)] + b_1 = a_1 t + d_1,$$

$$x_2(t) = a_2[t - t_{ср}(2)] + b_2 = a_2 t + d_2,$$

где

$$t_{ср}(1) = \frac{t_{11} + t_{21} + ... + t_{n(1)1}}{n(1)},$$
$$t_{ср}(2) = \frac{t_{12} + t_{22} + ... + t_{n(2)2}}{n(2)}. \qquad (1)$$

Таким образом,

$$d_k = b_k - a_k t_{ср}(k),\ k = 1, 2.$$

Преимущество модели с центрированием состоит в том, что асимптотическое описание совместного распределения коэффициентов проще в случае центрированной зависимости, например оценки коэффициентов a_k^*, b_k^*, $k = 1, 2$ асимптотически независимы.

Как известно, оценки метода наименьших квадратов имеют вид [2]

$$a_k^* = \frac{\sum_{i=1}^{n(k)} x_{ik}[t_{ik} - t_{ср}(k)]}{\sum_{i=1}^{n(k)} [t_{ik} - t_{ср}(k)]^2}, \qquad (2)$$

$$b_k^* = x_{ср}(k) = \frac{x_{1k} + x_{2k} + ... + x_{n(k)k}}{n(k)},\ k = 1, 2. \qquad (3)$$

Точечные оценки момента встречи $t_в^*$, уровня качества в момент встречи x^* и временно́го лага L^* выражаются через оценки коэффициентов линейных зависимостей следующим образом:

$$t_в^* = \frac{d_2^* - d_1^*}{a_1^* - a_2^*} = \frac{b_2^* - b_1^* + a_1^* t_{ср}(1) - a_2^* t_{ср}(2)}{a_1^* - a_2^*}, \qquad (4)$$

$$x^* = \frac{a_1^* d_2^* - a_2^* d_1^*}{a_1^* - a_2^*} =$$

$$= \frac{a_1^* b_2^* - a_2^* b_1^* + a_1^* a_2^* [t_{ср}(1) - t_{ср}(2)]}{a_1^* - a_2^*}, \qquad (5)$$

$$L^* = \frac{(a_2^* - a_1^*) t_0 + d_2^* - d_1^*}{a_2^*} =$$

$$= \frac{(a_2^* - a_1^*) t_0 + b_2^* - b_1^* + a_1^* t_{ср}(1) - a_2^* t_{ср}(2)}{a_2^*}. \qquad (6)$$

Из приведенных формул следует, что

$$t_в^* = f_1(a_1^*, a_2^*, b_1^*, b_2^*),\ x^* = f_2(a_1^*, a_2^*, b_1^*, b_2^*),$$

$$L^* = f_1(a_1^*, a_2^*, b_1^*, b_2^*),$$

где

$$f_1(z_1, z_2, z_3, z_4) = \frac{z_4 - z_3 + z_1 t_{ср}(1) - z_2 t_{ср}(2)}{z_1 - z_2};$$

$$f_2(z_1, z_2, z_3, z_4) =$$

$$= \frac{z_1 z_4 - z_2 z_3 + z_1 z_2 [t_{ср}(1) - t_{ср}(2)]}{z_1 - z_2};$$

$$f_3(z_1, z_2, z_3, z_4) =$$

$$= \frac{(z_2 - z_1) t_0 + z_4 - z_3 + z_1 t_{ср}(1) - z_2 t_{ср}(2)}{z_2}.$$

Поскольку все входящие в формулы моменты времени предполагаются заданными (детерминированными), то интересующие нас оценки задаются гладкими функциями от четырехмерного вектора $(a_1^*, a_2^*, b_1^*, b_2^*)$ оценок метода наименьших квадратов коэффициентов в линейных зависимостях.

Рассмотрим асимптотическое распределение вектора оценок МНК в рамках описанной выше непараметрической вероятностно-статистической модели [2]. Оценки $a_1^*, a_2^*, b_1^*, b_2^*$ являются несмещенными, их дисперсии

$$D(a_k^*) = \frac{\sigma_k^2}{\sum_{i=1}^{n(k)} [t_{ik} - t_{ср}(k)]^2},\ D(b_k^2) = \frac{\sigma_k^2}{n(k)},\ k = 1, 2.$$

Отметим, что в соответствии с принятыми предположениями все четыре дисперсии стремятся к нулю при безграничном росте $n(1)$ и $n(2)$.

Все ковариации вектора $(a_1^*, a_2^*, b_1^*, b_2^*)$ равны нулю. Для пар координат с различающимися нижними индексами это следует из предположения о независимости между собой совокупностей невязок, соответствующих измерениям значений двух разных линейных функций. Для пар координат с одинаковыми нижними индексами (т.е. для пар a_1^*, b_1^* и a_2^*, b_2^*) это установлено в работе [2]. Таким образом, в ковариационной матрице векторы $(a_1^*, a_2^*, b_1^*, b_2^*)$ отличны от нуля только элементы, стоящие на главной диагонали, т.е. дисперсии.

Каждый из векторов (a_1^*, b_1^*) и (a_2^*, b_2^*) является суммой $n(1)$ и $n(2)$ слагаемых соответственно. Если каждое из слагаемых мало по сравнению со всей суммой, т.е.

$$\lim_{n(k)\to\infty} \max_{1 \le i \le n(k)} \frac{|t_{ik} - t_{ср}(k)|}{\sqrt{\sum_{i=1}^{n(k)} [t_{ik} - t_{ср}(k)]^2}} = 0,\ k = 1, 2,$$

то при больших $n(1)$ и $n(2)$ распределение вектора $(a_1^*, a_2^*, b_1^*, b_2^*)$ с достаточной для практики точностью может быть заменено распределением нормального случайного вектора с независимыми координатами. Математические ожидания и дисперсии координат приближающего вектора совпадают с одноименными характеристиками вектора $(a_1^*, a_2^*, b_1^*, b_2^*)$. Другими словами, вектор $(a_1^*, a_2^*, b_1^*, b_2^*)$ является асимптотически нормальным с указанными выше параметрами.

Рассмотрим распределение функции от вектора оценок МНК. Если функция $f(z_1, z_2, z_3, z_4)$ достаточно гладкая, то согласно методу линеаризации [2, п. 4.4]

$$f(a_1^*, a_2^*, b_1^*, b_2^*) - f(a_1, a_2, b_1, b_2) = \frac{\partial f}{\partial z_1}(a_1^* - a_1) +$$

$$+ \frac{\partial f}{\partial z_2}(a_2^* - a_2) + \frac{\partial f}{\partial z_3}(b_1^* - b_1) + \frac{\partial f}{\partial z_4}(b_2^* - b_2)$$

с точностью до бесконечно малых величин более высокого порядка.

Как показано выше, правая часть последней формулы с достаточной для практики точностью может быть заменена суммой четырех независимых нормально распределенных величин с нулевыми математическими ожиданиями. Следовательно, функция $f(a_1^*, a_2^*, b_1^*, b_2^*)$ от вектора оценок МНК является асимптотически нормальной случайной величиной с математическим ожиданием $f(a_1, a_2, b_1, b_2)$, совпадающим с теоретическим значением, и дисперсией

$$Df(a_1^*, a_2^*, b_1^*, b_2^*) = \left(\frac{\partial f}{\partial z_1}\right)^2 D(a_1^*) +$$

$$+ \left(\frac{\partial f}{\partial z_2}\right)^2 D(a_2^*) + \left(\frac{\partial f}{\partial z_2}\right)^2 D(b_1^*) + \left(\frac{\partial f}{\partial z_4}\right)^2 D(b_2^*).$$

Подставив в это выражение приведенные выше значения дисперсий, получим

$$Df(a_1^*, a_2^*, b_1^*, b_2^*) = \left(\frac{\partial f}{\partial z_1}\right)^2 \frac{\sigma_1^2}{\sum_{i=1}^{n(1)} [t_{i1} - t_{ср}(1)]^2} +$$

$$+ \left(\frac{\partial f}{\partial z_2}\right)^2 \frac{\sigma_2^2}{\sum_{i=1}^{n(2)} [t_{i2} - t_{ср}(2)]^2} + \left(\frac{\partial f}{\partial z_2}\right)^2 \frac{\sigma_1^2}{n(1)} + \left(\frac{\partial f}{\partial z_4}\right)^2 \frac{\sigma_2^2}{n(2)}.$$

Применим разработанный подход к рассматриваемой задаче. Рассмотрим асимптотическое распределение момента встречи. Начнем с функции $t_в^* = f_1(a_1^*, a_2^*, b_1^*, b_2^*)$. В соответствии с общим подходом найдем частные производные:

$$\frac{\partial f_1}{\partial z_1} = \frac{z_3 - z_4 + z_2[t_{ср}(2) - t_{ср}(1)]}{(z_1 - z_2)^2},$$

$$\frac{\partial f_1}{\partial z_2} = \frac{z_4 - z_3 + z_1[t_{ср}(1) - t_{ср}(2)]}{(z_1 - z_2)^2},$$

$$\frac{\partial f_1}{\partial z_3} = -\frac{1}{z_1 - z_2}, \quad \frac{\partial f_1}{\partial z_4} = \frac{1}{z_1 - z_2}.$$

В приведенных выше формулах частные производные можно брать в точках как (a_1, a_2, b_1, b_2), так и $(a_1^*, a_2^*, b_1^*, b_2^*)$. Различие — бесконечно малые величины более высокого порядка. Поскольку истинные значения коэффициентов линейных зависимостей неизвестны, частные производные будем брать в точке $(a_1^*, a_2^*, b_1^*, b_2^*)$.

Из последних формул с помощью несложных преобразований получим

$$D(t_в^*) = \left\{\frac{b_1^* - b_2^* + a_2^*[t_{ср}(2) - t_{ср}(1)]}{(a_1^* - a_2^*)^2}\right\}^2 \frac{\sigma_1^2}{\sum_{i=1}^{n(1)}[t_{i1} - t_{ср}(1)]^2} +$$

$$+\left\{\frac{b_1^* - b_2^* + a_1^*[t_{ср}(2) - t_{ср}(1)]}{(a_1^* - a_2^*)^2}\right\}^2 \frac{\sigma_2^2}{\sum_{i=1}^{n(2)}[t_{i2} - t_{ср}(2)]^2} +$$

$$+\frac{1}{(a_1^* - a_2^*)^2}\left[\frac{\sigma_1^2}{n(1)} + \frac{\sigma_2^2}{n(2)}\right]. \qquad (7)$$

Для практического применения полученных результатов остается заменить неизвестные дисперсии невязок σ_1^2 и σ_2^2 на их состоятельные оценки. При больших объемах данных $n(1)$ и $n(2)$ используют оценки дисперсий невязок

$$(\sigma_1^2)^* = \frac{SS(1)}{n(1)}, \ (\sigma_2^2)^* = \frac{SS(2)}{n(2)}, \qquad (8)$$

где $SS(1)$ и $SS(2)$ — соответствующие остаточные суммы квадратов:

$$SS(k) = \sum_{i=1}^{n(k)}[x_{ik} - x_k^*(t_{ik})]^2, \ k = 1, 2. \qquad (9)$$

Иногда рекомендуют применение несмещенных оценок дисперсий невязок

$$(\sigma_1^2)^{**} = \frac{SS(1)}{n(1) - 2}, \ (\sigma_2^2)^{**} = \frac{SS(2)}{n(2) - 2}. \qquad (10)$$

Ясно, что с ростом объемов данных $n(1)$ и $n(2)$ различие между формулами (8) и (10) исчезает.

На основе полученных результатов легко указать методы доверительного оценивания и проверки гипотез для момента встречи $t_в$. Так, асимптотический доверительный интервал, соответствующий доверительной вероятности p, имеет вид

$$\left[t_в^* - U(p)\sqrt{D^*(t_в^*)}; t_в^* + U(p)\sqrt{D^*(t_в^*)}\right]. \qquad (11)$$

Здесь $D^*(t_в^*)$ — описанная выше оценка дисперсии случайной величины $t_в$ (с использованием той или иной оценки дисперсий невязок); $U(p)$ — квантиль стандартного нормального распределения порядка $(1 + p)/2$, т.е. $U(p) = \Phi^{-1}\left(\frac{1+p}{2}\right)$, где $\Phi(W)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1; $\Phi^{-1}(y)$ — обратная к ней функция.

В качестве примера рассмотрим показатели технического уровня продукции x_{i1} и x_{i2} (в условных единицах) двух предприятий — "Альфа" и "Бета" соответственно. Приведенные в таблице данные показывают, что в 1999 г. первое предприятие отстает от второго, но постепенно сокращает разрыв, более быстрыми темпами наращивая показатель x_{i1} технического уровня. Когда же оно догонит второе предприятие?

Показатели технического уровня продукции двух предприятий (в условных единицах, на конец года)

Показатель	Годы/условные моменты времени $t_{i1} = t_{i2}$						
	1999/1	2000/2	2001/3	2002/4	2003/5	2004/6	2005/7
x_{i1}	0,1	0,2	0,6	0,5	0,8	0,9	1,3
x_{i1}^*	0,06	0,25	0,44	0,63	0,82	1,01	1,2
x_{i2}	0,6	0,95	0,8	1,2	1,1	1,2	1,4
x_{i2}^*	0,71	0,82	0,93	1,04	1,15	1,26	1,37

Для проведения расчетов введем условные моменты времени (см. таблицу). Методом наименьших квадратов восстановим линейные независимости. По формуле (1) получим $t_{ср}(1) = t_{ср}(2) = 4$, а по формулам (2) и (3) — оценки коэффициентов линейных зависимостей

$$a_1^* = 0{,}19; \ a_2^* = 0{,}11; \ b_1^* = 0{,}63; \ b_2^* = 1{,}04.$$

Восстановленные зависимости примут вид

$$x_1^*(t) = 0{,}19(t-4) + 0{,}63,\ \ x_2^*(t) = 0{,}11(t-4) + 1{,}04.$$

Их восстановленные значения $x_{i1}^* = x_1^*(t_{i1})$ и $x_{i2}^* = x_2^*(t_{i2})$ приведены в таблице.

Момент встречи определим по формуле (4):

$$t_в^* = \frac{1{,}04 - 0{,}63 + 0{,}19 \cdot 4 - 0{,}11 \cdot 4}{0{,}19 - 0{,}11} = 9{,}13.$$

Другими словами, значения показателей технического уровня предприятий сравняются в начале 2008 г. Оценку уровня качества в момент встречи получим по формуле (5):

$$x^* = \frac{0{,}19 \cdot 0{,}6 - 0{,}11 \cdot (-0{,}13) + 0{,}19 \cdot 0{,}11(4-4)}{0{,}19 - 0{,}11} = 1{,}6.$$

По формуле (6) вычислим временно́й лаг, т.е. величину, показывающую, на сколько предприятие "Альфа" отстает от конкурента, например, в 2004 г.:

$$L^* = \frac{(0{,}11 - 0{,}19) \cdot 6 + 0{,}6 - (-0{,}13)}{0{,}11} = 2{,}27.$$

Рассчитаем по формуле (11) асимптотический доверительный интервал, соответствующий доверительной вероятности $p = 0{,}95$. Для этого значения несмещенных оценок дисперсий невязок найдем по формуле (10), а остаточной суммы квадратов $SS(1)$ и $SS(2)$ — по формуле (9): $\sigma_1^2 = 0{,}01382$, $\sigma_2^2 = 0{,}0157$. Асимптотическую дисперсию момента встречи определим по формуле (7): $D(t_в^*) = 4{,}379$. Поскольку $U(p) = 1{,}96$ при $p = 0{,}95$, то доверительный интервал имеет вид

$$\left[9{,}13 - 1{,}96\sqrt{4{,}379};\ 9{,}13 + 1{,}96\sqrt{4{,}379}\right] = [5{,}028;\ 13{,}23].$$

Таким образом, возможно, что обгон уже состоялся (в 2004 или в 2005 г.), но это не отражено в таблице из-за погрешностей, искажающих зависимости.

Необходимость получения оценок момента встречи $t_в^*$, уровня качества в момент встречи x^* и временно́го лага L^* была выявлена в результате решения прикладных проблем, возникших при разработке системы автоматического проектирования (САПР) стандартов на продукцию [3]. В этой системе реализуются функции информационного обеспечения и анализа данных о характеристиках качества группы отечественных и зарубежных образцов (марок, моделей) аналогичной продукции, а также о требованиях нормативно-технической документации на эту продукцию и поддерживаются функции принятия решений по управлению качеством, сертификации и стандартизации.

Статистические методы в САПР стандартов используются для анализа распределений показателей качества продукции, исследования взаимосвязей этих показателей, выявления групп продукции по уровню качества, анализа временны́х рядов и прогнозирования качества продукции. Основные сложности программной реализации этих методов связаны с проблемами решения задач пользователями (инженерами по стандартизации и техническому регулированию), не имеющими специальной подготовки по статистическим методам. Кроме того, возникают специфические задачи, требующие совершенствования известного статистического аппарата. К ним, в частности, относится задача сравнительного анализа тенденций развития отечественной и зарубежной групп продукции, решению которой и посвящена данная работа. Разработанное математическое и программное обеспечение применялось для анализа данных о характеристиках качества изделий электронной техники.

Описанные методы могут быть использованы при решении различных практических задач, связанных с интервальной оценкой точки пересечения двух регрессионных прямых.

Литература

1. *Robinson D. E.* / J. Am. Statist. Ass. 1964. V. 19. № 2. P. 214 – 238.
2. *Орлов А. И.* Прикладная статистика. — М.: Экзамен, 2006. — 672 с.
3. *Медведев В. Н., Орлов А. И.* — В сб.: Тезисы докладов III Всесоюзной школы-семинара "Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа". — М.: ЦЭМИ АН СССР, 1987. С. 313 – 314.