ТОЧНЫЕ D-ОПТИМАЛЬНЫЕ ПЛАНЫ ЭКСПЕРИМЕНТОВ ДЛЯ ЛИНИИ РЕГРЕССИИ РОБАСТНЫЕ ОТНОСИТЕЛЬНО ЛИНЕЙНОГО ВОЗМУЩЕНИЯ ДИСПЕРСИИ НАБЛЮДЕНИЙ

В. П. Кирлица
Белорусский государственный университет,
Минск, Беларусь
E-mail: Kirlitsa@bsu.by

Установлено, что точные *D*-оптимальные планы экспериментов остаются робастными относительно определенного класса линейных возмущений дисперсии наблюдений.

Ключевые слова: оптимальные планы экспериментов, линия регрессии.

Рассмотрим линейную модель наблюдений

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon(x_i), i = 1, \ldots, n, \tag{1}$$

где y_i – наблюдаемые переменные, θ_o, θ_1 – неизвестные параметры, x_i – контролируемые переменные из интервала [–1, 1], $\varepsilon(x_i)$ –не контролируемые и не наблюдаемые случайные ошибки наблюдений со средним значением равным нулю и дисперсией $D\{\varepsilon(x_i)\}$ = $d(x_i)$ изменяющейся кусочно-линейно

$$d(x) = d_1, x \in [-1, c], -1 \le c \le 1, d_1 > 0,$$
$$d(x) = \frac{(d_1 - d_2)x + d_2 c - d_1}{c - 1}, x \in [c, 1], d_2 > 0. \tag{2}$$

Из (2) следует, что на интервале [-1, *c*] наблюдения равноточные, а на интервале [*c*, 1] дисперсия наблюдений подвергается линейному возмущению.

При c = 1 и $d_1 > 0$ модель наблюдений (1), (2) обращается в хорошо исследованную модель равноточных наблюдений [1], для которой точный *D*-оптимальный план экспериментов имеет вид

$$\varepsilon_n^o = \begin{Bmatrix} -1, & 1 \\ n_1, & n - n_1 \end{Bmatrix}, \tag{3}$$

где $n_1 = s$ для четного числа наблюдений $n = 2s$, а для нечетных $n = 2s + 1$ имеем $n_1 = s$ либо $n_1 = s + 1$. Оценки неизвестных параметров, построенных по плану (3) не зависят от значения дисперсии d_1 и имеют вид

$$\hat{\theta}_0 = \frac{1}{2n_1(n - n_1)} \left[(n - n_1) \sum_{i=1}^{n_1} y_i + n_1 \sum_{i=1}^{n-n_1} y'_i \right], \tag{4}$$

$$\hat{\theta}_1 = \frac{1}{2n_1(n - n_1)} \left[n_1 \sum_{i=1}^{n-n_1} y'_i - (n - n_1) \sum_{i=1}^{n_1} y_i \right], \tag{5}$$

где y_i – наблюдения в точке –1, а y'_i – наблюдения в точке 1.

Ниже будет показано, что планы (3) и соответствующие им оценки (4), (5) для равноточных наблюдений останутся робастными, неизменными для определенных наборов параметров d_1, d_2, c, определяющих дисперсию (2) неравноточных наблюдений.

Теорема. Точки спектра точного D-оптимального плана экспериментов для линии регрессии (1) с дисперсией наблюдений (2) могут находиться лишь в точках –1, c, 1.

Доказательство. Пусть точки x_i^0, $i = 1,\ldots,n$, образуют точный D-оптимальный план. Допустим, что некоторая точка, например x_1^0, точного D-оптимального плана не совпадает ни с одной из точек –1, c, 1, т.е. принадлежит либо интервалу (–1, c), либо интервалу (c, 1). На каждом из этих интервалов дисперсия наблюдений $d(x)$ изменяется линейно, т.е. $d(x) = ax + b$, $x \in (x_-, x_+)$, где $a = 0$, $b = d_1$, $x_- = -1$, $x_+ = c$, если $x \in (-1, c)$, либо $a = (d_1 - d_2)/(c-1)$, $b = (d_2 c - d_1)/(c-1)$, $x_- = c$, $x_+ = 1$, если $x \in (c, 1)$. Точку x_1^0 сделаем “плавающей” на (x_-, x_+), т.е. заменим x_1^0 на x.

Исследуем поведение определителя информационной матрицы нового плана как функции аргумента x. Определитель информационной матрицы $M(x)$ нового плана равен:

$$|M(x)| = \frac{ex^2 - 2gx + f}{d(x)} + ef - g^2,$$

где

$$e = \sum_{i=2}^{n} \frac{1}{d(x_i^0)} > 0,\ f = \sum_{i=2}^{n} \frac{(x_i^0)^2}{d(x_i^0)},\ g = \sum_{i=2}^{n} \frac{x_i^0}{d(x_i^0)}.$$

Производная $|M(x)|$ равна:

$$\frac{aex^2 + 2bex - 2bg - af}{(ax+b)^2}. \tag{6}$$

Обозначим через $D = 4e(b^2 + 2abg + a^2 f)$ дискриминант квадратного трехчлена, стоящего в числителе формулы (6).

Если $D \le 0$, то производная (6) на интервале $[x_-, x_+]$ не меняет своего знака, т.е. $|M(x)|$ монотонно возрастает, либо убывает.

Если $D > 0$, то $|M(x)|$ – выпуклая функция на $[x_-, x_+]$. Действительно, так как $e > 0$, то

$$\frac{d^2|M(x)|}{dx^2} = \frac{2(b^2 e + 2abg + a^2 f)}{(ax+b)^3} > 0, x \in [x_-, x_+].$$

Итак, в любом случае функция $|M(x)|$ на $[x_-, x_+]$ достигает максимального значения на концах интервала $[x_-, x_+]$, что противоречит тому, что точка $x_1^0 \in (x_-, x_+)$ и является точкой спектра точного D-оптимального плана экспериментов. Теорема доказана.

Из теоремы следует, что точный D-оптимальный план ε_n^0 в своем спектре может содержать лишь точки –1, c, 1 и имеет вид:

$$\varepsilon_n^0 = \begin{Bmatrix} -1, & c & 1 \\ n_1^0, & n_2^0, & n - n_1^0 - n_2^0 \end{Bmatrix}, \tag{7}$$

где n_1^0, n_2^0 – решение задачи целочисленного квадратичного программирования:

$$f(n_1, n_2) = \frac{1}{kd_1^2}\{-4n_1^2 - (c-1)^2 n_2^2 + b(k,c)n_1 n_2 + 4nn_2 + n(c-1)^2 n_2\} \to \max, \tag{8}$$

где $b(k, c) = k(1 + c)^2 - c^2 + 2c - 5$, $k = \dfrac{d_2}{d_1}$, а максимум вычисляется по множеству

$$G = \{n_1, n_2 : 0 \le n_1 \le n-1, 0 \le n_2 \le n-1, 1 \le n_1 + n_2 \le n\}. \tag{9}$$

Задачу целочисленного квадратичного программирования (8), (9) можно решать прямым перебором значений функции $f(n_1, n_2)$ в узлах решетки множества G, число которых $\dfrac{n(n+1)}{2} - 3$. Однако от прямого перебора можно отказаться и решить задачу максимизации рациональнее, используя тот факт, что при каждом фиксированном значении $n_2 = s$, $s = 0, 1, \ldots, n-1$ функция $f(n_1, s)$, как функция аргумента n_1 – это парабола с ветвями направленными вниз.

Максимум функции $f(n_1, s)$ по n_1, без ограничений (9), достигается в точке, где ее производная обращается в ноль, т.е. в точке $n_1 = n_1(s)$:

$$n_1(s) = \frac{4n + b(k, c)\cdot s}{8}. \tag{10}$$

Максимум функции $f(n_1, s)$ на множестве G, при фиксированном s, зависит от того, будет ли точка (10) принадлежать интервалу $[g_1(s), g_2(s)]$ или нет. Здесь $g_1(s)$, $g_2(s)$ – нижняя и верхняя границы изменения n_1 на множестве G при фиксированном s. Очевидно, что $g_1(0) = 1$, $g_1(s) = 0$, $s \ge 1$, $g_2(0) = n - 1$, $g_2(s) = n - s$, $s \ge 1$.

Если $n_1(s) \le g_1(s)$, то максимум $f(n_1, s)$ по n_1 на множестве G, при фиксированном s, достигается в точке $n_1^0(s) = g_1(s)$.

Если $g_1(s) < n_1(s) < g_2(s)$, то для определения точки $n_1^0(s)$, в которой будет достигаться максимум $f(n_1, s)$ по n_1 на множестве G при фиксированном s, значение $n_1(s)$ надо округлить до ближайшего целого значения, т.е. $n_1^0(s) = [n_1(s)]$, если $\{n_1(s)\} < 0.5$. Здесь и далее символы $[z]$, $\{z\}$ означают, соответственно, целую и дробную часть числа z. Если $\{n_1(s)\} = 0.5$, то $n_1^0(s) = [n_1(s)]$ либо $n_1^0(s) = [n_1(s)] + 1$. Наконец, если $\{n_1(s)\} > 0.5$, то $n_1^0(s) = [n_1(s)] + 1$.

Если $n_1(s) \ge n - s$, то $n_1^0(s) = n - s$.

Затем, вычисляем значения функций $f(n_1^0, s) = f^0(s)$. Последовательно сравнивая значения $f^0(s)$, $s = 0, 1, \ldots, n-1$ друг с другом, начиная с первого значения $f^0(0)$, находим решение задачи целочисленного программирования (8), (9).

Результаты численного решения задачи (8), (9) для некоторых значений параметров k, c, n при $d_1 = 1$ приведены в таблице 1.

Таблица 1

k	c	n	n_1^0	n_2^0	$\lvert M(\varepsilon_n^0)\rvert$
3	0.9	10	5	5	90.25
3	0.5	4	2	2	9
100	- 0.9	4	1	1	0.1622
100	- 0.8	4	2	1	0.1924
3	0.5	100	50	50	5625

Для некоторых наборов параметров d_1, d_2, c, определяющих дисперсию (2), для построения D-оптимального плана экспериментов необязательно решать задачу целочисленного программирования (8), (9), можно получить качественные результаты.

Итак, согласно [3], построим параболу $\varphi(x)$, проходящую через точки $(-1, d_1)$, $(1, d_2)$:

$$\varphi(x) = \frac{d_1}{4}\{(1+k)x^2 + 2(k-1)x + 1 + k\}, k = \frac{d_2}{d_1}.$$

В статье [3] обосновано, что если

$$d(x) \geq \varphi(x), x \in [-1,1], \tag{11}$$

то точный D-оптимальный план ε_n^0 для неравноточных наблюдений (2) определяется формулой (3), т.е. совпадает с оптимальным планом для равноточных наблюдений. Поэтому важно установить, в каких случаях будет выполняться неравенство (11). Можно выделить три таких случая.

1) Для $0 < k \leq 1$ неравенство (11) выполняется. Это следует из того, что $\varphi(x)$ – это выпуклая функция, а $d(x) \geq (1-\lambda)d_1 + \lambda d_2$, $0 \leq \lambda \leq 1$. Итак, можно утверждать, что для неравноточных наблюдений, описываемых дисперсией (2), для которой $d_1 \geq d_2 > 0$, точный D-оптимальный план совпадает с планом (3) для равноточных наблюдений. Другими словами, классический D-оптимальный план (3) для равноточных наблюдений остается робастным относительно изменения дисперсии наблюдений, описываемых (2) с $d_1 \geq d_2 > 0$. Оценки параметров для D-оптимального плана с неравноточными наблюдениями (2) имеют точно такой же вид (4), (5), как и для равноточных наблюдений, т.е. робастны относительно возмущения дисперсии наблюдений (2) при $0 < k \leq 1$. Последнее утверждение следует из того, что в процессе построения оценок параметров дисперсии d_1, d_2, участвующие в построении оценок, взаимно сокращаются.

2) Для случая, когда $c = -1$, т.е. дисперсия наблюдений линейно возрастает либо убывает на интервале $[-1, 1]$, неравенство (11) также выполняется, Это непосредственно следует из того, что $\varphi(x)$ – это выпуклая функция. Следовательно, те выводы, которые были получены в пункте 1) относительно D-оптимального плана ε_n^0 и оценок неизвестных параметров, остаются в силе и в данном случае.

3) Для $d_2 > d_1$, $k > 1$, $-1 < c < 1$, неравенство (11) будет иметь место, если $-1 < c \leq c_1$, где $c_1 = (3-k) / (1+k)$. Здесь c_1 – корень уравнения $\varphi(x) = d_1$. Действительно, -1 и c_1 – корни уравнения $\varphi(x) = d_1$ и поэтому парабола $\varphi(x) \leq d_1$ для $-1 \leq x \leq c$. То, что (11) выполняется для $c \leq x \leq 1$ следует из того, что функция $\varphi(x)$ выпуклая. Следовательно, как и в пункте 1), D-оптимальный план (3) и соответствующие ему оценки параметров (4), (5) будут робастны относительно возмущения дисперсии (2) для $d_2 > d_1$ и $-1 < c \leq c_1$.

3.1) Для $d_2 > d_1$ и $c = c_1$ для нечетного числа наблюдений $n = 2m+1$ точный D-оптимальный план для неравноточных наблюдений, наряду с (3), может иметь особую структуру

$$\varepsilon_n^0 = \begin{Bmatrix} -1, & c_{1,} & 1 \\ m, & 1, & m \end{Bmatrix}, \tag{12}$$

а соответствующие ему оценки параметров имеют вид:

$$\hat{\theta}_0 = \frac{1}{4m}\left[\sum_{i=1}^{m} y_i + y_{c_1} + \sum_{i=1}^{m} y_i'\right], \tag{13}$$

$$\hat{\theta}_1 = \frac{1}{4m}\left[\sum_{i=1}^{m} y_i' - \sum_{i=1}^{m} y_i - y_{c_1}\right], \tag{14}$$

где y_i и y_i' имеют прежний смысл, а y_{c_1} – наблюдение в точке c_1. Из (13), (14) видно, что они не зависят от d_1, d_2, т.е. их можно строить так, как будто план (12) имеет равноточные наблюдения.

Особую структуру, отличную от классической, имеют точные D-оптимальные планы и соответствующие им оценки неизвестных параметров, если дисперсия наблюдений $d(x)$ линейно возрастает от нуля до d_2 на интервале [–1, 1], когда $c = -1$, $d_1 = 0$. В этом случае в точке –1 наблюдения могут проводиться точно, без ошибок. Проведем одно наблюдение в точке –1:

$$y_{-1} = \theta_0 - \theta_1, \tag{15}$$

где y_{-1} – наблюдение в точке –1. Из (15) следует, что

$$\theta_0 = y_{-1} + \theta_1. \tag{16}$$

Соотношения (15), (16) выполняются почти наверное, т.е. с вероятностью единица. Более одного наблюдения в точке –1 не имеет смысла проводить, так как будут получаться одинаковые соотношения (15), которые не будут нести дополнительной информации. Остальные n – 1 наблюдений, с учетом (16), нужно проводить согласно модели наблюдений

$$\bar{y}_i = \theta_1 t_i + \varepsilon(t_i - 1), i = 1, 2, \cdots, n-1, \tag{17}$$

с одним неизвестным параметром θ_1, при этом

$$D\{\varepsilon(t_i - 1)\} = d(t_i) = \frac{d_2 t_i}{2}.$$

В модели наблюдений(17)

$$\bar{y}_i = y_i - y_{-1}, 1 + x_i = t_i, t_i \in (0,2].$$

Для модели наблюдений (17) информационная матрица M вырождается в число

$$M = \sum_{i=1}^{n-1} \frac{t_i^2}{d(t_i)} = \frac{2}{d_2} \sum_{i=1}^{n-1} t_i.$$

Максимум этого числа достигается тогда, когда все $t_i = 2$, т.е. когда $x_i = 1$, $i = 1, 2, \ldots, n - 1$. В этом случае точный D-оптимальный план равен

$$\varepsilon_n^0 = \begin{Bmatrix} -1, & 1 \\ 1, & n-1 \end{Bmatrix}, \tag{18}$$

а соответствующие ему оценки неизвестных параметров таковы:

$$\hat{\theta}_1 = \frac{1}{2(N-1)} \sum_{I=1}^{N-1} Y_I - \frac{1}{2} y_{-1}, \hat{\theta}_0 = y_{-1} + \hat{\theta}_1. \tag{19}$$

В заключении осталось рассмотреть случай, когда дисперсия наблюдений $d(x) = 0$ на интервале [–1, c], $-1 < c < 1$, а на интервале [c, 1] она линейно возрастает до d_2. В этом случае можно получить, почти наверное, точные значения параметров θ_0, θ_1, если провести два наблюдения в различных точках из интервала [–1, c].

ЛИТЕРАТУРА

1. *Moyssiadis C., Kounias C.* Exact D-optimal N Observations 2^k Designs of Resolutions 3, when $N \equiv 1$ *mod* 4 / C. Moysiadis, C. Kounias // Math. Operationsforsch. U. Statist. 1983. № 3. P. 367 – 379.

2. *Федоров В.В.* Теория оптимального эксперимента / В.В. Федоров. М: Наука, 1968.

3. *Кирлица В.П.* *D*-оптимальные планы экспериментов, робастные относительно изменения дисперсии наблюдений / В.П. Кирлица // Вестн. БГУ. Сер. 1. 2008. № 3. С. 89 – 92.