

**Методические указания к  
индивидуальным заданиям по курсу  
«Планирование эксперимента в  
экономике»**

**А. И. Корякин**

Курс: Планирование эксперимента в экономике  
Семестр 9, 2009 год  
[portal.tpu.ru](http://portal.tpu.ru)

Учебная дисциплина «Планирование эксперимента в экономике» читается студентам специальности «Математические методы в экономике» с целью усвоения теоретических знаний о статистической теории оптимального планирования эксперимента и обработки его результатов, приобретения практических навыков применения этой теории в прикладных экономических исследованиях. Это позволит научиться не только правильно обрабатывать уже имеющуюся экономическую информацию, но и планировать сбор данных так, чтобы минимизировать затраты на получение достаточно точной экономической информации. Обучение этой дисциплине базируется на сумме знаний и умений, полученных студентами при изучении курсов «Теория вероятностей», «Математическая статистика», «Экономико-математическое моделирование», «Информационные технологии в экономике» и других.

Содержание курса «Планирование эксперимента в экономике»: экономические задачи планирования и обработки эксперимента; экстремальная задача выбора оценки и плана эксперимента; оценки по методу наименьших квадратов, наилучшая линейная оценка, простейшая монте-карловская оценка, оценка по методу наименьших квадратов со случайными узлами, их дисперсионные матрицы и области применения; критерии эффективности и теоремы эквивалентности; аналитические способы и численные алгоритмы построения оптимальных планов; моделирование непрерывных планов для оценок со случайными узлами. Индивидуальные домашние задания выполняются слушателями курса «Планирование эксперимента в экономике» с целью практического освоения методики и приобретения навыков использования методов и алгоритмов оптимального планирования и обработки эксперимента, оценивания точности полученных результатов для решения прикладных экономических задач.

Предполагается, что в некоторых  $n$ -мерных точках  $x^j = (x_1^j, \dots, x_n^j) \in X \subset \mathbb{R}_n$ ,  $j = 1, \dots, N$ , области возможных измерений  $X$  в результате эксперимента могут быть получены наблюдения  $\zeta(x^j, \omega^j)$  функции  $f(x^j)$ . Эти наблюдения содержат случайные ошибки, такие что

$$M_\omega \zeta(x^j, \omega^j) = f(x^j), \quad M_\omega [\zeta(x^j, \omega^j) \zeta(x^q, \omega^q)] = f(x^j) f(x^q) + \sigma^2(x^j, x^q)$$

и  $\sigma^2$  интегрируема по обоим аргументам в  $X$ . По результатам таких наблюдений необходимо оценить функциональную зависимость  $y = f(x)$ , о которой известно лишь, что

$$f(x) = \sum_{i=1}^m c_i \varphi_i(x) + r(m, x),$$

где  $\varphi_i(x)$ ,  $i = 1, \dots, m$ , — заданный набор линейно независимых на  $X$  функций. Функция  $g(x, m) = \sum_{i=1}^m \theta_i \varphi_i(x)$

будет проекционной оценкой (регрессией) функции  $f(x)$ , если  $\theta_i$ ,  $i = 1, \dots, m$ , — оценки коэффициентов разложения  $c_i$ , полученные по наблюдениям  $\zeta(x^j, \omega^j)$ ,  $j = 1, \dots, N$ .

Задача состоит в выборе таких оценок  $\theta_i$  и таких узлов эксперимента  $x^j$ ,  $j = 1, \dots, N$ , которые бы обеспечивали наиболее точную по вероятности проекционную оценку функциональной зависимости  $y = f(x)$ .

Введем матричные обозначения

$$\begin{aligned} Z_{N,1} &= (\zeta(x^1, \omega^1) \dots \zeta(x^N, \omega^N))^T; \\ F_{N,1} &= (f(x^1) \dots f(x^N))^T; \\ R_{N,1} &= (r(m, x^1) \dots r(m, x^N))^T; \quad M_\omega Z = F; \\ M_\omega [ZZ^T] - FF^T &= \Sigma_{N,N}, \end{aligned}$$

где  $\Sigma$  — матрица ковариаций ошибок наблюдений с элементами  $\sigma^2(x^j, x^q)$ ,  $j, q = 1, \dots, N$ ;

$$\begin{aligned} \Phi_{m,1}(x) &= (\varphi_1(x) \dots \varphi_m(x))^T; \quad A_{N,m} = (\Phi(x^1) \dots \Phi(x^N))^T; \\ c_{m,1} &= (c_1 \dots c_m)^T; \quad F = Ac + R. \end{aligned}$$

Теорема Гаусса–Маркова дает так называемую наилучшую линейную оценку  $\hat{c}$  коэффициентов  $c$  и соответствующую ей проекционную оценку  $\hat{c}^T \Phi(x)$  функции  $f(x)$ . Эта оценка по вероятности оптимальна как в пространстве коэффициентов, так и в пространстве функций на множестве всех оценок вида  $TZ$ , для которых  $TA = E$ . В случае независимости ошибок наблюдений (матрица  $\Sigma$  диагональна) и точно заданной модели ( $R = \bar{0}$ ) наилучшая линейная оценка имеет вид

$$\hat{c} = (A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} Z = \left[ \sum_{j=1}^N \frac{\Phi(x^j) \Phi^T(x^j)}{\sigma^2(x^j)} \right]^{-1} \sum_{j=1}^N \frac{\Phi(x^j) \zeta(x^j, \omega^j)}{\sigma^2(x^j)}.$$

(предполагается, что точки  $x^j$ ,  $j = 1, \dots, N \geq m$ , выбраны так, чтобы существовали обратные матрицы.) При этом матрица ковариаций ошибок такой оценки

$$K \hat{c} = \mathcal{M}[(c - \hat{c})(c - \hat{c})^T] = (A^T \Sigma^{-1} A)^{-1},$$

дисперсия оценки регрессии

$$d(x) = \mathcal{M}[f(x) - \hat{c}^T \Phi(x)]^2 = \Phi^T(x) K \hat{c} \Phi(x)$$

и  $\hat{c}^T \Phi(x) \pm \sqrt{d(x)}/\varepsilon$  — коридор ошибок оценивания функции  $f(x)$  с уровнем доверия  $1 - \varepsilon^2$ .

Оценка  $\hat{c}$  реализуется, если дисперсия ошибок наблюдений известна хотя бы с точностью до множителя, т.е.

$\sigma^2(x) = \sigma^2/\lambda(x)$ , где  $\sigma^2$  — неизвестная константа, а  $\lambda(x)$  — известная функция эффективности эксперимента. Тогда

$$K\hat{c} = \sigma^2 \left[ \sum_{j=1}^N \Phi(x^j)\Phi^T(x^j)\lambda(x^j) \right]^{-1}$$

и неизвестную  $\sigma^2$  можно оценить по формуле

$$\hat{\sigma}^2 = \frac{1}{N-m} \sum_{j=1}^N [\zeta(x^j, \omega^j) - \hat{c}^T \Phi(x^j)]^2.$$

Это позволяет (используя неравенство Чебышева) получить представление о доверительных интервалах наблюдений  $\zeta(x^j, \omega^j)$ , ошибках оценок  $\hat{c}_i$ ; коэффициентов разложения  $c_i$ , о коридоре ошибок проекционной оценки  $\hat{c}^T \Phi(x)$  функции  $f(x)$ .

Набор узлов  $x^j$ ,  $j = 1, \dots, N$ , в которых получают наблюдения, называют *планом эксперимента*. В плане может быть всего  $T \leq N$  разных узлов  $x^q$ ,  $q = 1, \dots, T$ , и в каждом по  $K_q$

наблюдений,  $\sum_{q=1}^T K_q = N$ . Тогда план задается в виде

$\varepsilon(N) = \{(x^1, K_1), \dots, (x^T, K_T)\}$  или в нормированном виде  $\tilde{\varepsilon}(N) = \{(x^1, p_1), \dots, (x^T, p_T)\}$ , где  $p_q = K_q/N$  — частота попадания в точку  $x^q$ . План  $\varepsilon$  называют непрерывным, если вероятности попадания  $p_q$  полагают непрерывными.

Непрерывный план может быть реализован, только если все  $p_q N$  целые.

Чтобы выделить в явном виде влияние величин  $\sigma^2$  и  $N$  на ошибки оценок, вводят нормированную информационную матрицу

$$I_n(\varepsilon) = \frac{\sigma^2}{N} [K\hat{c}(\varepsilon)]^{-1} = \sum_{q=1}^T p_q \lambda(x^q) \Phi(x^q) \Phi^T(x^q)$$

и нормированную дисперсию

$$d_n(x, \varepsilon) = \Phi^T(x) I_n^{-1}(\varepsilon) \Phi(x) = \frac{d(x, \varepsilon)}{\sigma^2} N.$$

Планы  $\varepsilon$ , максимизирующие  $\det I_n(\varepsilon)$ , называют *D-оптимальными*; минимизирующие  $\text{Sp } I_n(\varepsilon)$  называют *A-оптимальными*; минимизирующие  $\max_{x \in X} d_n(x, \varepsilon)$  — *минимаксными*, а минимизирующие среднее на  $X$  значение  $d_n(x, \varepsilon)$  — *Q-оптимальными*. Оптимальные планы известны лишь для достаточно простых моделей  $f(x) = \sum_{i=1}^m c_i \varphi_i(x)$  и областей  $X$ .

Итерационный алгоритм, сходящийся к  $D$ -оптимальному плану, основан на одной из теорем эквивалентности. Она утверждает, что план  $\varepsilon^*$   $D$ -оптимален тогда и только тогда, когда

$$\min_{\varepsilon} \max_{x \in X} d_H(x, \varepsilon) \lambda(x) = \max_{x \in X} d_H(x, \varepsilon^*) \lambda(x) = m.$$

Поэтому план  $\varepsilon$  можно улучшить, если добавить в него точку  $\tilde{x}$ , в которой

$$\max_{x \in X} d_H(x, \varepsilon) \lambda(x) = d_H(\tilde{x}, \varepsilon) \lambda(\tilde{x}) > m.$$

1. Пусть задан невырожденный и не  $D$ -оптимальный план  $\varepsilon_0 = \{(x^1, p_1), \dots, (x^{T_0}, p_{T_0})\}$ ,  $T_0 > m$ , пусть  $s = 0$ .

2. Находим  $I_H(\varepsilon_s)$ ,  $d_H(x, \varepsilon_s)$ ,

$\max_{x \in X} d_H(x, \varepsilon_s) \lambda(x) = d_H(\tilde{x}_s, \varepsilon_s) \lambda(\tilde{x}_s) = H(\varepsilon_s)$ , т.е. получаем план

$\varepsilon(\tilde{x}_s) = \{(\tilde{x}_s, 1)\}$  из одной точки.

3. Составляем план

$$\begin{aligned} \varepsilon_{s+1} &= \varepsilon_s(1 - \alpha_s^*) + \alpha_s^* \varepsilon(\tilde{x}_s) = \\ &= \{(x^1, (1 - \alpha_s^*)p_1), \dots, (x^{T_s}, (1 - \alpha_s^*)p_{T_s}), (x^{T_s+1}, \alpha_s^*)\}, \end{aligned}$$

где  $x^{T_s+1} = \tilde{x}_s$ ;  $\alpha_s^* = [H(\varepsilon_s) - m]/[mH(\varepsilon_s) - m]$ ;  $T_s + 1 = T_{s+1}$ .

4. Увеличиваем номер итерации  $s$  на единицу и возвращаемся к пункту 2 с новым планом  $\varepsilon_{s+1}$  вместо  $\varepsilon_s$ .

Итерационный процесс продолжается, пока величина  $H(\varepsilon_s)$  не достигнет достаточно близкого к  $m$  значения. Тогда непрерывный план  $\varepsilon_s$  можно считать «почти»  $D$ -оптимальным.

При реализации этого алгоритма рекомендуется на каждой итерации выдавать значения  $I_H(\varepsilon_s)$ ,  $\det I_H(\varepsilon_s)$ ,  $d_H(x, \varepsilon_s)$ ,  $\tilde{x}_s$ ,  $H(\varepsilon_s)$ ,  $\alpha_s^*$  и план  $\varepsilon_{s+1}$ , уделяя внимание точности обращения матриц и нахождения точек  $\tilde{x}_s$ . Может оказаться, что проще решать уравнения  $\max_{x \in X} d_H(x, \varepsilon_s)\lambda(x) = d_H(\tilde{x}_s, \varepsilon_s)\lambda(\tilde{x}_s)$

приближенно, например, вычисляя значения  $d_H(x, \varepsilon_s)\lambda(x)$  в узлах достаточно мелкой сетки, покрывающей область возможных измерений  $X$ .

Если точкой  $\tilde{x}_s$  наибольшего значения  $d_H(x, \varepsilon_s)\lambda(x)$  окажется точка, уже входящая в план  $\varepsilon_s$  с весом  $p_q$ , значит, в план  $\varepsilon_{s+1}$  эта точка войдет один раз с весом  $\alpha_s^* + (1 - \alpha_s^*)p_q$ . Может оказаться, что наибольшее значение  $d_H(x, \varepsilon_s)\lambda(x)$  получится в нескольких точках области  $X$ . Тогда рекомендуется включать в новый план  $\varepsilon_{s+1}$  сразу все эти точки с одинаковыми весами, сумма которых равна  $\alpha_s^*$ .

Если итерационный процесс сходится медленно, а закономерность появления узлов и изменения весов последовательности планов  $\varepsilon_s$  простая, то можно ускорить сходимость, экстраполировав процесс изменения планов  $\varepsilon_s$  с ростом числа итераций.

При точно заданном количестве наблюдений  $N$  реализовать оптимальный непрерывный план  $\varepsilon^*$  удастся лишь приближенно. Дискретный план  $\tilde{\varepsilon}(N)$  можно получить из непрерывного  $\varepsilon^* = \{(x^1, p_1), \dots, (x^T, p_T)\}$ , моделируя с помощью датчика случайных чисел  $N$  случайных  $n$ -мерных величин, принимающих дискретные значения  $x^q$ ,  $q = 1, \dots, T$ , с вероятностями  $p_q$ . Можно получить дискретный план и другим способом. Следует распределить  $N_0 < N$  наблюдений, взяв количество наблюдений  $K_q$  в точках  $x^q$  равным целой части  $p_q N$ ,  $q = 1, \dots, T$ . Оставшиеся  $N - N_0 < T$  наблюдений добавить по одному к тем  $K_q$ , где дробная часть  $p_q N$  наибольшая.

Необходимо проверить качество полученного таким образом дискретного плана  $\tilde{\varepsilon}(N)$ , вычислив  $\det I_n(x, \tilde{\varepsilon}(N))$ ,  $d_n(x, \tilde{\varepsilon}(N))$  и  $H(\tilde{\varepsilon}(N))$ .