АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ
№ 2 1974

КРАТКИЕ СООБЩЕНИЯ

УДК 519.2

В. Ф. ДОДУЛ
(Москва)

ОЦЕНКА ОБЪЕМА ВЫБОРКИ ИЗМЕРЕНИЙ, НЕОБХОДИМОГО ДЛЯ ОПРЕДЕЛЕНИЯ УРАВНЕНИЯ КВАДРАТИЧЕСКОЙ РЕГРЕССИИ С ЗАДАННОЙ ТОЧНОСТЬЮ

Одна из важных задач планирования регрессионных экспериментов оценка объема выборки измерений переменных X и Y, необходимого для определения регрессионной зависимости между ними с требуемой точностью. В [1] эта задача решалась применительно к линейной регрессии. В настоящей работе искомым является уравнение квадратической регрессии.

Как известно, исходной информацией, используемой при планировании экспериментов по определению регрессионной зависимости, являются моменты распределения зависимой и независимой переменных. Практически вид закона распределения независимой величины, а также моменты этого закона определяются априори, исходя из теоретических и физических соображений. При этом с достаточными основаниями закон распределения результатов измерения независимой переменной X полагается нормальным.

Законы же совместного распределения зависимой и независимой переменных, а также самой зависимой переменной и моменты этих распределений, как правило, априори неизвестны, и для их определения наряду с использованием априорных данных необходима постановка специальных предварительных экспериментов.

Очевидно, что в случае нелинейной регрессионной зависимости между переменными X и Y их совместное распределение отличается от нормального. Вместе с тем известно, что среди распределений, обладающих одинаковыми математическими ожиданиями и ковариационными матрицами случайных величин, нормальное распределение содержит минимум информации о величинах, распределенных по этому закону. Поэтому выражения для объема выборки n, полученные с учетом предположения о нормальном законе совместного распределения переменных X и Y, связанных нелинейной зависимостью, дадут верхнюю границу требуемого объема выборки. При этом требуемая точность оценки искомой регрессионной зависимости будет заведомо выполняться.

Так как истинный закон совместного распределения переменных X и Y не известен, то в настоящей работе будем исходить из наиболее неблагоприятного случая нормального совместного распределения переменных X и Y. Это предположение в значительной степени оправданно для больших выборок ($n>30$), но при малых объемах выборок ($n<20$) для определения требуемого объема выборки необходимо знание истинного закона совместного распределения переменных X и Y.

Постановка задачи следующая. Пусть зависимая переменная Y связана с независимой переменной X регрессионной зависимостью

$$y=ax^2+bx+c. \tag{1}$$

Последовательность независимых измерений переменной X распределена по нормальному закону. Совместное распределение значений переменных X и Y полагается нормальным.

Математическое ожидание m_x, дисперсия D_x независимой переменной X и дисперсия D_{x^2} значений X^2 известны. Оценки $\hat{D}_y$ дисперсии зависимой переменной и коэффициентов корреляции $\hat{r}_{xy}$ и $\hat{r}_{x^2y}$ известны с заданной достоверностью, определяемой доверительными вероятностями β и доверительными интервалами $I_{\hat{D}_y}$, $I_{\hat{r}_{xy}}$ и $I_{\hat{r}_{x^2y}}$ соответственно.

Требуется опеределить объем выборки n, необходимый для определения регрессионной зависимости (1) с заданной точностью. Перейдем к решению поставленной задачи.

Ввиду того, что априорная информация о поведении искомой зависимости (1) отсутствует, с целью получения такой информации производятся предварительные измерения с объемом выборки n_0. Неизвестные коэффициенты a, b и c, определяющие поведение линии регрессии $y(x)$, являются функциями выборочных значений переменных x_i и y_i

($i=1, 2, \ldots, n_0$). Для их определения воспользуемся принципом наименьших квадратов, согласно которому значения a, b и c находятся из условия

$$\sum_{i=1}^{n_0} \left[y_i - \left(ax_i^2 + bx_i + c\right)\right]^2 = \min.$$

Продифференцируем сумму квадратов $\sum_{i=1}^{n_0} \left[y_i - \left(ax_i^2 + bx_i + c\right)\right]^2$ по a, b и c и, приравняв производные нулю, находим:

$$\sum_{i=1}^{n_0} y_i - \left(ax_i^2 + bx_i + c\right)\Big] x_i^2 = 0;$$

$$\sum_{i=1}^{n_0} \left[y_i - \left(ax_i^2 + bx_i + c\right)\right] x_i = 0;$$

$$\sum_{i=1}^{n_0} \left[y_i - \left(ax_i^2 + bx_i + c\right)\right] = 0.$$

Раскрывая скобки, производя суммирование и деление на n_0, получим систему линейных уравнений

$$\begin{cases} \alpha_4 a + \alpha_3 b + \alpha_2 c = \hat{\alpha}_{2,1}; \\ \alpha_3 a + \alpha_2 b + \alpha_1 c = \hat{\alpha}_{1,1}; \\ \alpha_2 a + \alpha_1 b + \alpha_0 c = \hat{\alpha}_{0,1}, \end{cases} \tag{2}$$

где α_0, α_1, α_2, α_3, α_4 — моменты распределения независимой переменной X;$\hat{\alpha}_{0,1}$, $\hat{\alpha}_{1,1}$, $\hat{\alpha}_{2,1}$ — оценки моментов совместного распределения зависимой и независимой переменных по результатам предварительных измерений. Следуя [2], нетрудно показать, что моменты нормального распределения имеют вид:

$$\begin{aligned} &\alpha_0 = 1; \\ &\alpha_1 = m_x; \\ &\alpha_2 = m_x^2 + D_x; \\ &\alpha_3 = 3m_x D_x + m_x^3; \\ &\alpha_4 = m_x^4 + 6m_x^2 D_x + 3D_x^2; \\ &\hat{\alpha}_{0,1} = \hat{m}_y; \\ &\hat{\alpha}_{1,1} = \hat{K}_{xy}\frac{n-1}{n} + m_x\hat{m}_y; \\ &\hat{\alpha}_{2,1} = \hat{K}_{x^2y}\frac{n-1}{n} + 2m_x\hat{K}_{xy}\frac{n-1}{n} + \hat{m}_y D_x + m_x^2\hat{m}_y, \end{aligned} \tag{3}$$

где $\hat{K}_{xy}$ и $\hat{K}_{x^2y}$ — оценки для корреляционных моментов совместного распределения переменных X и Y по результатам предварительных измерений. С учетом соотношений (3) находим следующее решение системы уравнений (2):

$$\begin{aligned} &a = \frac{\hat{K}_{x^2y}}{2D_x^2}\frac{n-1}{n}; \\ &b = \left(\frac{\hat{K}_{xy}}{D_x} - \frac{m_x\hat{K}_{x^2y}}{D_x^2}\right)\frac{n-1}{n}; \\ &c = \hat{m}_y - \left(\frac{m_x\hat{K}_{xy}}{D_x} + \frac{m_x^2\hat{K}_{x^2y}}{2D_x^2} - \frac{\hat{K}_{x^2y}}{2D_x}\right)\frac{n-1}{n}. \end{aligned} \tag{4}$$

В соответствии с [2] степень отклонения искомой регрессионной зависимости от реально существующей будем характеризовать дисперсией предсказания линии регрессии

$$\overline{D}_y = \frac{1}{n-l}\sum_{i=1}^{n} \left[y_i - \left(ax_i^2 + bx_i + \quad\right)\right]^2, \tag{5}$$

где n — общий объем измерений; l — число неизвестных коэффициентов. Подставив в (5) значения коэффициентов регрессии (4), получим

$$\overline{D}_y = \frac{n}{n-3}\hat{D}_y \left\{1 + \left[\frac{m_x^4 \hat{r}_{x^2y}^2 D_{x^2}}{D_x^4} - 2\frac{m_x^2 \hat{r}_{x^2y}^2 D_{x^2}}{D_x^3} + \frac{\hat{r}_{x^2y}^2 D_{x^2}}{2D_x^2} - \hat{r}_{xy}^2\right]\left(\frac{n-1}{n}\right)\right\}, \quad (6)$$

откуда уравнение для требуемого объема выборки измерений переменных X и Y будет иметь вид

$$\left[\overline{D}_y - \hat{D}_y(1+R)\right] n^2 - \hat{D}_y(3-2R)\,n - \hat{D}_y R = 0, \quad (7)$$

где

$$R = \frac{m_x^4 \hat{r}_{x^2y}^2 D_{x^2}}{D_x^4} - 2\frac{m_x^2 \hat{r}_{x^2y}^2 D_{x^2}}{D_x^3} + \frac{\hat{r}_{x^2y}^2 D_{x^2}}{2D_x^2} - \hat{r}_{xy}^2 .$$

Искомый объем выборки является решением квадратичного уравнения (7), удовлетворяющим условию $n \geqslant 3$.

В соответствии с [3] при большом n_0 и нормальном законе распределения измерений Y доверительный интервал для оценки дисперсии $\hat{D}_y$ определяется по приближенной формуле

$$I_{\hat{D}_y} = \frac{t_\beta D_y}{n_0} \approx \frac{t_\beta \hat{D}_y}{n_0},$$

где D_y — теоретическое значение дисперсии. Величина t_β находится из уравнения $\Phi(t_\beta) = \beta$, где $\Phi(t_\beta)$ — функция Лапласа. Доверительно интервалы для оценок коэффициентов корреляции $\hat{r}_{xy}$ и $\hat{r}_{x^2y}$ получаем по приближенной формуле

$$I_{\hat{r}} \approx 2t_\beta \sqrt{D_{\hat{r}}},$$

где в соответствии с [3]

$$\sqrt{D_{\hat{r}}} = \frac{1-r^2}{\sqrt{n_0}} \approx \frac{1-\hat{r}^2}{\sqrt{n_0}} .$$

Приближенная замена теоретического значения коэффициента r его оценкой $\hat{r}$ допустима, если n_0 значительно, а r не очень близко к ± 1.

При решении (7) в качестве значения дисперсии D_y используется ее верхняя доверительная граница, равная $\hat{D}_y + \frac{t_\beta \hat{D}_y}{2n_0}$, а в качестве значений корреляционных коэффициентов r_{xy} и r_{x^2y} — их нижние доверительные границы, равные $\hat{r} - t_\beta \sqrt{D_{\hat{r}}}$.

При таком использовании исходных данных, а также предположении о нормальном законе совместного распределения переменных X и Y выражение (7) дает оценку сверху для требуемого объема выборки. Анализ влияния различных факторов на требуемый объем выборки измерений позволяет сделать следующие выводы:

1) объем выборки не зависит от среднего значения зависимой переменной m_y;

2) объем выборки возрастает с увеличением дисперсий D_x, D_{x^2} и D_y; особенно резкое увеличение объема выборки необходимо при возрастании отношения m_x/D_x;

3) уменьшение объема выборки происходит с увеличением коэффициентов корреляции r_{xy} и r_{x^2y} (особенно для больших значений r_{x^2y}), а также со снижением требования к точности оценки линии регрессии.

Результаты настоящей работы могут быть использованы для планирования регрессионных экспериментов, связанных с оптимизацией технологических процессов, оценкой характеристик сложных систем и изучением механизма физических процессов.

Однако они справедливы лишь для больших выборок ($n > 30$). Если же требуемый для обеспечения заданной точности определения регрессионной зависимости (1) объем выборки невелик ($n < 20$), то его величина нуждается в уточнении.

ЛИТЕРАТУРА

1. Ю. А. Зак. Необходимое число наблюдений для получения уравнений линейной регрессии с заданной точностью.— Сборник трудов украинского НИИ целлюлозно-бумажной промышленности, вып. 8. Киев, 1967.
2. Е. И. Пустыльник. Статистические методы анализа и обработки наблюдений. М., «Наука», 1968.
3. Я. И. Лукомский. Теория корреляции и ее применение к анализу производства. М., Госстатиздат, 1961.

Поступило в редакцию 12 мая 1970 г.