Метод существенной выборки для оценивания границ доверительных интервалов в задачах параметрической нелинейной регрессии

Горлова Марина Владимировна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: д.ф.-м.н., профессор М.С. Ермаков
Рецензент: к.ф.-м.н., доцент Ю.Н. Каштанов

Санкт-Петербург

Введение

Редкие события находят применение во многих приложениях и задачах математической статистики.

При вычислении вероятности редких событий прямым моделированием возникает две проблемы:

- мощные вычислительные ресурсы
- малейшие ошибки в моделировании влекут значительные ошибки оценок

Наиболее распространенный метод вычисления – ***метод существенной выборки***.

Введение

Редкие события находят применение во многих приложениях и задачах математической статистики.

При вычислении вероятности редких событий прямым моделированием возникает две проблемы:

- мощные вычислительные ресурсы
- малейшие ошибки в моделировании влекут значительные ошибки оценок

Наиболее распространенный метод вычисления – ***метод существенной выборки***.

Введение

Редкие события находят применение во многих приложениях и задачах математической статистики.

При вычислении вероятности редких событий прямым моделированием возникает две проблемы:

- мощные вычислительные ресурсы
- малейшие ошибки в моделировании влекут значительные ошибки оценок

Наиболее распространенный метод вычисления – ***метод существенной выборки***.

Пусть $b_n > 0$, $b_n \to 0$, $nb_n \to \infty$ при $n \to \infty$.
Задача — оценить вероятность

$$V_n = P(T(\hat{P}_n) - T(P_0) > b_n), \tag{1}$$

где P_0 – теоретическое распределение, $\hat{P}_n$ – эмпирическая фунция распределения, построенная по наблюдениям x_i с распределением P_0, $1 \leqslant i \leqslant n$, T – некоторый функционал.

Метод существенной выборки

- Введем меру P_n: $P_n \ll P_0$. Обозначим $q_n = \dfrac{dP_n}{dP_0}$.
- Промоделируем k независимых выборок с распределением P_n
$$Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}, \quad 1 \leqslant i \leqslant k.$$
- В качестве оценки вероятности (1) берем
$$\hat{V}_n = \frac{1}{k} \sum_{i=1}^{k} \chi(T(\hat{P}_n^{(i)}) - T(P_0) > b_n) \prod_{j=1}^{n} q_n^{-1}(Y_j^{(i)}), \qquad (2)$$
где $\hat{P}_n^{(i)}$ эмпирическое распределение $Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}$.

Постановка задачи

Задана модель нелинейной регрессии

$$x_i = S(t_i, \theta) + \xi_i, \quad 1 \leqslant i \leqslant n, \tag{3}$$

где $S(t, \theta)$ – нелинейная функция, $t_1, \ldots, t_n$ – точки, равномерно взятые на отрезке $[0, 1]$, $\theta = (\theta^1, \ldots, \theta^l)$ – вектор неизвестных парамеров, ξ_i – независимые нормально распределенные случайные величины с плотностью распределения $f(x)$.

Истинное значение вектора пареметров θ_0, $\hat{\theta}_n$ – его оценка.

Цель работы – оценить вероятность

$$V_n = P((\hat{\theta}_n - \theta_0) > b_n). \tag{4}$$

Метод решения

Плотность распределения x_i в модели регрессии при $\theta = \theta_0$

$$p_{i,\theta_0}(x) = f(x - S(t_i, \theta_0)). \tag{5}$$

Замена меры

$$p_{i,\theta_n}(x) = f(x - S(t_i, \theta_0 + b_n)). \tag{6}$$

$$q_n(x) = \prod_{i=1}^{n} \frac{p_{i,\theta_0}(x)}{p_{i,\theta_n}(x)}. \tag{7}$$

Моделируем k независимых выборок, $Y_j^{(i)}$ с плотностью $p_{j,\theta_n}(x)$

$$Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}, \quad 1 \leqslant i \leqslant k.$$

Оценка V_n

$$\hat{V}_n = \frac{1}{k} \sum_{i=1}^{k} \chi(\hat{\theta}_n > \theta_0 + b_n) \prod_{j=1}^{n} q_n^{-1}(Y_j^{(i)}). \tag{8}$$

Метод решения

Плотность распределения x_i в модели регрессии при $\theta = \theta_0$

$$p_{i,\theta_0}(x) = f(x - S(t_i, \theta_0)). \tag{5}$$

Замена меры

$$p_{i,\theta_n}(x) = f(x - S(t_i, \theta_0 + b_n)). \tag{6}$$

$$q_n(x) = \prod_{i=1}^{n} \frac{p_{i,\theta_0}(x)}{p_{i,\theta_n}(x)}. \tag{7}$$

Моделируем k независимых выборок, $Y_j^{(i)}$ с плотностью $p_{j,\theta_n}(x)$

$$Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}, \quad 1 \leqslant i \leqslant k.$$

Оценка V_n

$$\hat{V}_n = \frac{1}{k} \sum_{i=1}^{k} \chi(\hat{\theta}_n > \theta_0 + b_n) \prod_{j=1}^{n} q_n^{-1}(Y_j^{(i)}). \tag{8}$$

Метод решения

Плотность распределения x_i в модели регрессии при $\theta = \theta_0$

$$p_{i,\theta_0}(x) = f(x - S(t_i, \theta_0)). \tag{5}$$

Замена меры

$$p_{i,\theta_n}(x) = f(x - S(t_i, \theta_0 + b_n)). \tag{6}$$

$$q_n(x) = \prod_{i=1}^{n} \frac{p_{i,\theta_0}(x)}{p_{i,\theta_n}(x)}. \tag{7}$$

Моделируем k независимых выборок, $Y_j^{(i)}$ с плотностью $p_{j,\theta_n}(x)$

$$Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}, \quad 1 \leqslant i \leqslant k.$$

Оценка V_n

$$\hat{V}_n = \frac{1}{k} \sum_{i=1}^{k} \chi(\hat{\theta}_n > \theta_0 + b_n) \prod_{j=1}^{n} q_n^{-1}(Y_j^{(i)}). \tag{8}$$

Метод решения

Плотность распределения x_i в модели регрессии при $\theta = \theta_0$

$$p_{i,\theta_0}(x) = f(x - S(t_i, \theta_0)). \tag{5}$$

Замена меры

$$p_{i,\theta_n}(x) = f(x - S(t_i, \theta_0 + b_n)). \tag{6}$$

$$q_n(x) = \prod_{i=1}^{n} \frac{p_{i,\theta_0}(x)}{p_{i,\theta_n}(x)}. \tag{7}$$

Моделируем k независимых выборок, $Y_j^{(i)}$ с плотностью $p_{j,\theta_n}(x)$

$$Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}, \quad 1 \leqslant i \leqslant k.$$

Оценка V_n

$$\hat{V}_n = \frac{1}{k} \sum_{i=1}^{k} \chi(\hat{\theta}_n > \theta_0 + b_n) \prod_{j=1}^{n} q_n^{-1}(Y_j^{(i)}). \tag{8}$$

Асимптотическая эффективность процедуры существенной выборки

Математическое ожидание оценки

$$\omega_n = \mathsf{E}\,\hat{V}_n = V_n.$$

Дисперсия оценки

$$\mathsf{Var}[\hat{V}_n] = U_n - \omega_n^2, \tag{9}$$

где

$$U_n = \mathsf{E}_p \left(\frac{1}{k} \sum_{i=1}^{k} \chi_{\{(\hat{\theta}_n^{(i)} - \theta) > b_n\}} \prod_{j=1}^{n} q_n^{-1}(Y_j^{(i)}) \right)^2 .$$

Асимптотическая эффективность процедуры существенной выборки

Введем определения

Определение

Процедура называется асимптотически эффективной (в смысле логарифмической асимптотики), если

$$\varlimsup_{n\to\infty} \frac{\log U_n}{2\log \omega_n} = 1.$$

Определение

Процедура называется эффективной, если

$$\varlimsup_{n\to\infty} \frac{U_n}{\omega_n^2} = 1.$$

Асимптотическая эффективность процедуры существенной выборки

Пусть для задачи (1) выполнено

1. Функция g принадлежит множеству Φ, где Φ – множество функций f, $\mathsf{E}\left[f(X)\right] = 0$, таких что
$$\lim_{n\to\infty} (nb_n^2)^{-1} \log[nP_0(|f(X)| > nb_n)] = -\infty.$$
2. Множество Λ_Φ всех мер $Q \in \Lambda$, таких что $\int\limits_\Omega |f|\,dQ < \infty \ \ \forall f \in \Phi$.
$\Lambda_{0\Phi}$ - множество всех зарядов $G = P - R$, где $P, R \in \Lambda$.
Существует полунорма $N \in \Lambda_{0\Phi}$, такая что $\forall Q \in \Lambda_\Phi$
$$|T(Q) - T(P_0) - \int\limits_\Omega g\,dQ| \leqslant$$
$$\omega(N(Q - P_0), \int\limits_\Omega g\,dQ, T(Q) - T(P_0))$$

Асимптотическая эффективность процедуры существенной выборки

с функцией $\omega : \mathsf{R}^3 \longrightarrow \mathsf{R}^1_+$, такой что

$$\lim_{t_1,t_2,t_3 \to 0} \frac{\omega(t_1,t_2,t_3)}{t_1+t_2+t_3} = 0.$$

Теорема (М.С. Ермаков, 2007)

В предположениях 1 и 2 рассмотрим процедуру существенной выборки, основанную на в. м. Q_n с плотностью $q_{1n}(x) = \lambda_n + b_n h(x)\chi\left(h(x) > -\delta b_n^{-1}\right)$ или $q_{2n}(x) = c_n exp\{b_n h(x)\}\cdot \cdot\chi\left(h(x) < \delta b_n^{-1}\right)$, где λ_n, c_n – константы нормализации, $0 < \delta < 1$ и $E(h(X)) = 0, E|h(X)| < \infty, E|h^2(X)| < \infty$, тогда процедура существенной выборки асимптотически эффективна, если $h = \sigma_g^{-2} g$.

Асимптотическая эффективность процедуры существенной выборки

Сформулируем основной результат в виде теоремы

Теорема

Пусть выполнены условия

1. $b_n > 0, b_n \to 0,\ nb_n^{2+\alpha} \to \infty$ *при* $n \to \infty, 1 \leqslant \alpha \leqslant 0.$
2. *sup*$_{t,\theta}|S(t,\theta)| < C.$
3. $S(t,\theta)$ $\forall t$ *и* $\forall\theta$ *дифференцируема по* θ *и верно* $|S(t,\theta+b_n) - S(t,\theta) - b_n S'_\theta(t,\theta)| < cb_n^{1+\frac{\alpha}{2}}.$

Тогда построенная процедура существенной выборки является асимптотически эффективной.

Первое условие означает, что рассматриваемая задача является задачей об умеренных уклонениях.

Численное моделирование

Проведем численное моделирование на примере модели

$$S(t,\theta) = 1 - \frac{2}{(1+\exp(1+\theta t))^2}. \qquad (10)$$

Модель описывает содержание биологически активных веществ в растворах.

Оценку θ будем вычислять методом максимального правдоподобия.

Изобразим полученные оценки при разных k на рисунке 1.

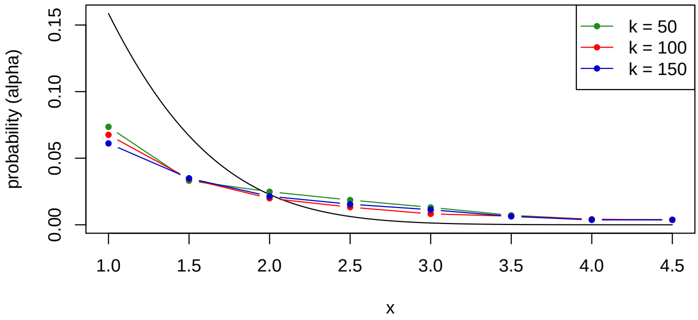

Рис. 1: Оценки вероятности при k = 50, 100, 150

На рисунке отобразим точность оценок 2.

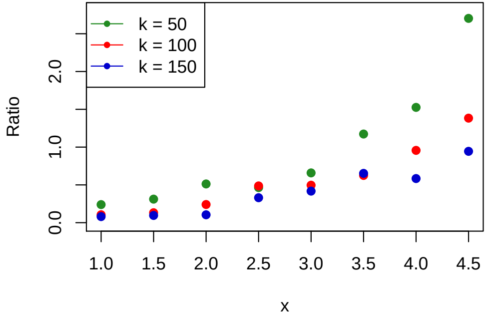

Рис. 2: Отношение дисперсии оценок к квадрату среднего при $k = 50, 100, 150$

Изобразим полученные оценки при разных n на рисунке 3.

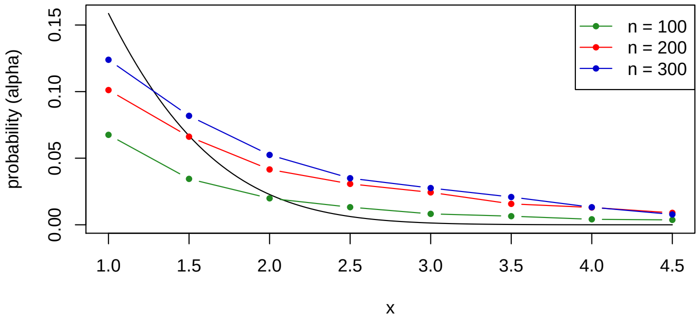

Рис. 3: Оценки вероятности при n = 100, 200, 300

Численное моделирование. Зависимость от n

Изобразим полученные эффективности оценок при $n = 100, 200, 300$ на рисунках 4 и 5.

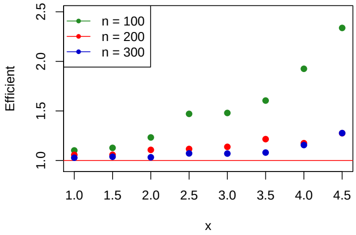

Рис. 4: Эффективность

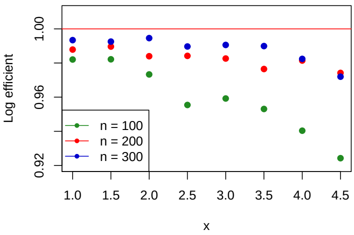

Рис. 5: Асимптотическая эффективность

Построим доверительные интервалы для $n = 100$ и $n = 300$ на рисунках 6 и 7.

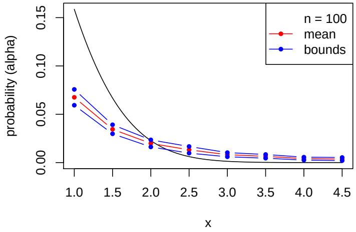

Рис. 6: Доверительные интервалы для $n = 100$

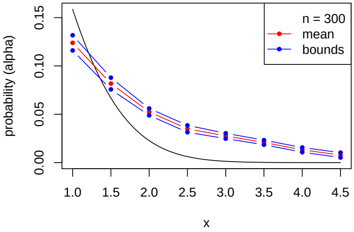

Рис. 7: Доверительные интервалы для $n = 300$

Заключение

Полученные результаты

1. Был применен метод вычисления редких событий (метод существенной выборки).
2. Доказана асимптотическая эффективность процедуры существенной выборки в зоне вероятностей умеренных уклонений.
3. Проведено численное моделирование, в результате которого построены доверительные интервалы для оценок вероятностей и сосчитаны их эффективности. Моделирование численно показало, что метод применим в зоне умеренных уклонений.

Спасибо за внимание!