Построение и статистический анализ некоторых моделей выбора

Сукманская Ксения Ивановна, гр. 522

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

Научный руководитель: к.ф.-м.н. Голяндина Н.Э.
Рецензент: к.ф.-м.н. Некруткин В.В.

Санкт-Петербург
2012г.

Модель двухэтапного опроса

Схема двухэтапного выбора:

Первоначальный выбор

из перечня предлагаемых вариантов

$\|$

некоторое

внешнее

воздействие

$\Downarrow$

Повторный выбор

из того же перечня

Интересует «сила» оказываемого внешнего воздействия, которое может иметь как положительный, так и отрицательный эффект.

m — количество продуктов, из них k рекламируемых.
Параметры:

- p_i, $i = 0, \dots, m-2$ — доли продуктов, $(p_{m-1} = 1 - \sum_{i=0}^{m-2} p_i)$,
- p_{ch} — вероятность хаотичного выбора,
- p_{adv} — вероятность воздействия рекламы на человека,
- $p^{(j)}$, $j = 0, \dots, k-2$ — условные вероятности влияния рекламы продукта под номером j, $(p^{(k-1)} = 1 - \sum_{i=0}^{k-2} p^{(i)})$.

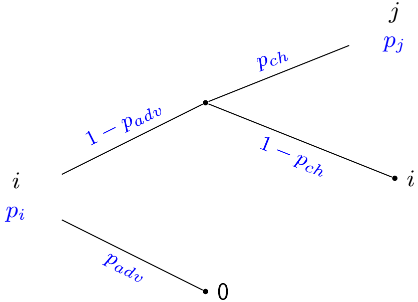
Параметрическое множество:
$\{\theta = (p_0, \dots, p_{m-2}, p_{ch}, p_{adv}, p^{(0)}, \dots, p^{(k-2)}) \in \Theta \subset (0,1)^{m+k}\}$.
Предполагаем, что истинное значение параметра θ^0 с некоторой окрестностью лежит в Θ.

Модель описывается видом вероятностей: $p_{ij} = p_{ij}(\theta)$ — вероятность выбрать i–тый продукт до рекламы и j–тый — после $(i, j = 0, \dots, m-1)$.

Построенные модели:

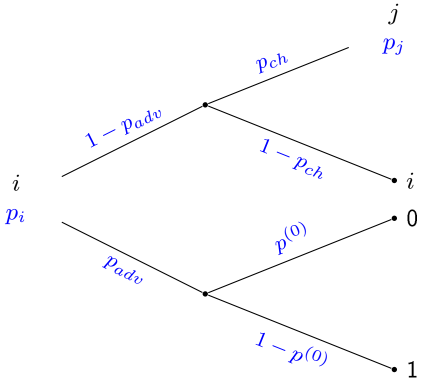
- Модель с положительным эффектом рекламы одного рекламируемого продукта.[Голяндина Н.Э.];
- Модель с отрицательным эффектом рекламы одного рекламируемого продукта;
- Модель с двумя рекламируемыми продуктами: положительный эффект реклам продуктов под номерами «0» и «1»;
- Модель с двумя рекламируемыми продуктами: отрицательный эффект рекламы продукта под номером «0» и положительный эффект рекламы продукта под номером «1».

Модель с положительным эффектом рекламы одного рекламируемого продукта.[Голяндина Н.Э.]



$$
\begin{aligned}
p_{00}(\theta) &= p_0(p_{adv} + (1 - p_{adv})(1 - p_{ch} + p_{ch}p_0)), \\
p_{i0}(\theta) &= p_i(p_{adv} + (1 - p_{adv})p_{ch}p_0), \quad i \neq 0, \\
p_{ii}(\theta) &= p_i(1 - p_{adv})(1 - p_{ch} + p_{ch}p_i), \quad i \neq 0, \\
p_{ij}(\theta) &= p_i(1 - p_{adv})p_{ch}p_j, \quad j \neq 0, i \neq j.
\end{aligned}
$$

Модель двухэтапного опроса с двумя рекламируемыми продуктами: положительный эффект реклам продуктов под номерами «0» и «1».



Задачи

План исследования модели:

- Построение простых оценок параметров с помощью гладкого отображения и получение вида их ковариационной матрицы;
- Улучшение простых оценок и вид их ковариационной матрицы;
- Сравнение дисперсий простых и улучшенных оценок.

Способ построения простых оценок

$p_{ij}(\theta)$ — вероятность выбора i–го продукта до рекламы и j–го — после.

$$
\begin{aligned}
p_{00}(\theta) &= p_0\big(p_{adv}p^{(0)} + (1-p_{adv})(1-p_{ch}+p_{ch}p_0)\big),\\
p_{11}(\theta) &= p_1\big(p_{adv}(1-p^{(0)}) + (1-p_{adv})(1-p_{ch}+p_{ch}p_1)\big),\\
p_{i0}(\theta) &= p_i\big(p_{adv}p^{(0)} + (1-p_{adv})p_{ch}p_j\big), \quad i \neq 0,\\
p_{i1}(\theta) &= p_i\big(p_{adv}(1-p^{(0)}) + (1-p_{adv})p_{ch}p_j\big), \quad i \neq 1,\\
p_{ij}(\theta) &= p_i(1-p_{adv})p_{ch}p_j, \quad i \neq j,\ j \neq 0,1,\\
p_{ii}(\theta) &= p_i(1-p_{adv})(1-p_{ch}+p_{ch}p_i), \quad i \neq 0,1.
\end{aligned}
$$

n_{ij} — количество людей, сменивших предпочтение с i–го продукта на j–й, n — количество опрашиваемых.

$$
\begin{aligned}
\frac{n_{00}}{n} &= p_0\Big(p_{adv}p^{(0)} + (1-p_{adv})(1-p_{ch}+p_{ch}p_0)\Big),\\
\frac{n_{11}}{n} &= p_1\Big(p_{adv}(1-p^{(0)}) + (1-p_{adv})(1-p_{ch}+p_{ch}p_1)\Big),\\
\frac{n_{i0}}{n} &= p_i\Big(p_{adv}p^{(0)} + (1-p_{adv})p_{ch}p_j\Big), \quad i \neq 0,\\
\frac{n_{i1}}{n} &= p_i\Big(p_{adv}(1-p^{(0)}) + (1-p_{adv})p_{ch}p_j\Big), \quad i \neq 1,\\
\frac{n_{ij}}{n} &= p_i(1-p_{adv})p_{ch}p_j, \quad i \neq j,\ j \neq 0, 1,\\
\frac{n_{ii}}{n} &= p_i(1-p_{adv})(1-p_{ch}+p_{ch}p_i), \quad i \neq 0, 1.
\end{aligned}
$$

$$\widetilde{\theta} = (\widetilde{p}_i,\ \widetilde{p}_{ch},\ \widetilde{p}_{adv},\ \widetilde{p}^{(0)})$$ — простые оценки $(i = 0, \ldots, m-2)$.

$$\begin{aligned}
\widetilde{p}_i &= n_{i\cdot}/n,\ i = 0, \ldots m-2, \\
\widetilde{p}_{ch} &= n \frac{\sum\limits_{\{i,j:i\neq j,\ i,j\neq 0,1\}} n_{ij}}{\sum\limits_{\{i,j:i\neq j,\ i,j\neq 0,1\}} (n_{i\cdot} n_{j\cdot})} \cdot \frac{n - (n_{0\cdot} + n_{1\cdot})}{n - (n_{\cdot 0} + n_{\cdot 1})}, \\
\widetilde{p}_{adv} &= \frac{n_{\cdot 0} + n_{\cdot 1} - (n_{0\cdot} + n_{1\cdot})}{n - (n_{0\cdot} + n_{1\cdot})}, \\
\widetilde{p}^{(0)} &= \frac{n_{\cdot 0} - (1 - \tilde{p}_{adv}) n_{0\cdot}}{n \tilde{p}_{adv}},
\end{aligned}$$

где $n_{i\cdot} = \sum_{j=0}^{m-1} n_{ij},\ n_{\cdot j} = \sum_{i=0}^{m-1} n_{ij}$.

Вид ковариационной матрицы простых оценок

- Для $\xi_n = (n_{00}/n, \ldots, n_{m-1,m-1}/n)^{\mathrm{T}}$ выполнено:

$$\mathcal{L}(\sqrt{n}(\xi_n - z_0)) \overset{n\to\infty}{\Longrightarrow} \mathcal{N}(\mathbf{0}, \Sigma),$$

где $z_0 = (p_{00}, \ldots, p_{m-1,m-1})^{\mathrm{T}}$,
$\Sigma = \| \delta_{\alpha\beta} p_\alpha - p_\alpha p_\beta \|$, где $\alpha,\ \beta = (i,j),\ (i,j = 0, \ldots, m-1)$.

- Простые оценки: $\widetilde{\theta} = H(\xi_n),\ H(z_0) = \theta^0$,

$H(z)$ — гладкое отображение на Θ, $\Delta H = \{\partial H_i(z)/\partial z^j\}_{i,j=0}^{m-1}$ тогда [Боровков Л.Л. Гл.5, §5, Т.3В]:

$$\mathcal{L}\left(\sqrt{n}(\widetilde{\theta} - \theta^0)\right) \overset{n\to\infty}{\Longrightarrow} \mathcal{N}(\mathbf{0}, \Sigma_H), \qquad \text{где } \Sigma_H = \Delta H^{\mathrm{T}}(z_0)\, \Sigma\, \Delta H(z_0).$$

Пусть $X = (X_1, \dots, X_n) \mapsto (n_1, \dots, n_s)$ — повторная выборка из распределения P_{θ^0}.
Логарифм функции правдоподобия:

$$l(\theta, X) = \sum_{k=1}^{s} n_k \ln p_k(\theta).$$

Предполагаем, что выполнены условия регулярности оценок максимального правдоподобия.
Информационная матрица:

$$I_{ij}(\theta) = \sum_{k=1}^{s} \frac{1}{p_k(\theta)} \frac{\partial p_k(\theta)}{\partial \theta_i} \frac{\partial p_k(\theta)}{\partial \theta_j}, \quad i, j = 1 \dots, r.$$

Определение

Асимптотически эквивалентными оценке максимального правдоподобия (а.э.о.м.п.) будем называть оценки $\hat{\theta}$, для которых выполнено

$$\mathcal{L}(\sqrt{n}(\hat{\theta} - \theta^0)) \Rightarrow \mathcal{N}(\boldsymbol{0}, I^{-1}(\theta^0)).$$

Одношаговые оценки:

$$\hat{\theta} = \bar{\theta} + \frac{1}{n} I^{-1}(\bar{\theta}) \cdot \dot{l}(\bar{\theta}). \tag{1}$$

Теорема (Закс. Ш., Т.5.5.4.)

Если начальные оценки $\bar{\theta}$ такие, что $n^{1/4}(\bar{\theta} - \theta^0) = o_p(1)$ при $n \to \infty$, то одношаговые оценки являются а.э.о.м.п.

Простые оценки $\widetilde{\theta}$:

- состоятельные,
- $\mathcal{L}\left(\sqrt{n}(\widetilde{\theta} - \theta^0)\right) \Rightarrow \mathcal{N}\left(\mathbf{0}, \Sigma_H\right).$

Следовательно, в (1) можно взять $\bar{\theta} = \widetilde{\theta}$.

Результаты: сравнения простых и одношаговых оценок

Результаты численного сравнения асимптотических дисперсий простых и одношаговых оценок:

- $\widetilde{p}_0$, $\widetilde{p}_1$ — являются а.э.о.м.п.;
- оценка $\widetilde{p}_{ch}$ «плохая», улучшение значительное: в среднем на $60 - 80\%$;
- при больших значениях p_{adv} оценки $\widetilde{p}_{adv}$ и $\widetilde{p}^{(0)}$ — «хорошие», с дисперсией близкой к минимальной,
  при малых значениях p_{adv} улучшение составляет $\approx 20\%$.

Реализация метода построения а.э.о.м.п.

Сложности:

- вычисление простых оценок по выборке;
- вычисление информационной матрицы и обратной к ней.

Решение:

- строковое задание формул для вероятностей $p_{ij}(\theta)$, начальных оценок $\widetilde{\theta}$ и реализация метода с помощью символьных вычислений.

Средства для решения:

- программа TSC[С.Кокорин, А.Демский] предоставляет возможность вычисления простых и одношаговых оценок для базовой модели с одним рекламируемым продуктом;
- модификация программы средствами ООП позволяет добавлять новые модели двухэтапного опроса с целью вычисления оценок параметров.

Применение к реальным данным

Данные: эксперимент двухэтапного опроса с одним рекламируемым продуктом.

категория продукта	$\hat{p}_0$	$\hat{p}_{ch}$	$\hat{p}_{adv}$	n
Подгузники	0.698	0.136	0.040	106
Стир. Порошок	0.137	0.261	0.033	284
Косметика	0.411	0.293	0.078	185

Таблица : Одношаговые оценки по базовой модели: «0» — объединение всех продуктов того же бренда, что и рекламируемый продукт.

категория продукта	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_{ch}$	$\hat{p}_{adv}$	$\hat{p}^{(0)}$	n
Подгузники	0.302	0.396	0.144	0.063	0.866	106
Стир. Порошок	0.099	0.039	0.263	0.039	0.699	284
Косметика	0.027	0.384	0.307	0.103	0.884	185

Таблица : Оценки параметров модели с двумя рекламами: «0» — действительно рекламируемый продукт, «1» — объединение остальных того же бренда.

Считаем, что p_{ch}, p_{adv} и $p^{(j)}$, $j = 0, \dots, k-2$ фиксированы.
Положим $\theta = (p_0, \dots, p_{m-1})$, $\tau \in (0,1)^m$ — вектор «предпочтений»,
тогда имеем: $p_{ij}(\theta, \tau) = \theta_i \; \pi_{ij}(\tau)$.

Обозначения:

- $\theta^{(n)} \in (0,1)^m$, $\sum\limits_{i=0}^{m-1} \theta_i^{(n)} = 1$, — распределение долей продуктов перед n–ой рекламой, $\theta^{(1)} = (p_0, \dots, p_{m-1})$;
- $\tau^{(n)} \in (0,1)^m$, $\sum\limits_{i=0}^{m-1} \tau_i^{(n)} = 1$ — «предпочтения» при случайном выборе после n–той рекламы, $\tau^{(1)} = \theta^{(1)}$;
- $\Pi(\tau^{(n)}):\quad p_{ij}(\theta^{(n)}, \tau^{(n)}) = \theta_i^{(n)} \; \pi_{ij}(\tau^{(n)})$.

Многоэтапная модель задает последовательность распределений $\theta^{(n)}$:

$$\theta^{(n+1)} = \theta^{(n)} \; \Pi(\tau^{(n)}).$$

Задача: $\lim\limits_{n\to\infty} \theta^{(n)} = \theta^* \ ?$

Две схемы принятия решения на каждом следующем этапе:

1. без корректировки предпочтений: $\tau^{(n)} = \tau^{(1)}$.

   Тогда θ^* ищем из уравнения:

$$\theta^* = \theta^* \ \Pi.$$

2. с корректировкой предпочтений: $\tau^{(n)} = \theta^{(n)}$.

   Тогда θ^* ищем из уравнения:

$$\theta^* = \theta^* \ \Pi(\theta^*).$$

Модель с двумя положительно–рекламируемыми продуктами

1. без корректировки предпочтений:

$$\theta_i^* = \frac{1}{G}\begin{cases} p_{adv}p^{(0)} + (1-p_{adv})p_{ch}p_0 & \text{при} \quad i=0,\\ p_{adv}(1-p^{(0)}) + (1-p_{adv})p_{ch}p_1 & \text{при} \quad i=1,\\ (1-p_{adv})p_{ch}p_i & \text{при} \quad i>1, \end{cases}$$

где $i = 2, \dots, m-1, \;\; G = p_{adv} + p_{ch}(1-p_{adv})$.

Свойства:

- $\partial\theta_j^*/\partial p_{ch} = -p_{adv}(1-p_{adv})(p^{(j)} - p_j)/G^2, \;\; j = 0, 1;$
- $\partial\theta_j^*/\partial p_{adv} = p_{ch}(p^{(j)} - p_j)/G^2, \;\; j = 0, 1;$
- $\partial\theta_0^*/\partial p^{(0)} = p_{adv}/G > 0, \quad \partial\theta_1^*/\partial p^{(0)} = -p_{adv}/G < 0.$

2. с корректировкой предпочтений: $\theta^* = (p^{(0)}, 1-p^{(0)}, 0\dots, 0)$.

Заключение

Основные результаты работы:

- Построены три модификации базовой модели двухэтапного опроса;
- Для каждой модели получены «простые» и а.э.о.м.п. оценки, проведено их сравнение;
- Модифицирована программа для автоматизации вычисления одношаговых оценок, с целью удобства добавления новых моделей. С ее помошью сделана обработка реальных данных и проведена интерпретация результатов;
- Исследованы и проинтерпретированы предельные распределения последовательности долей продуктов на рынке, при повторном воздействии рекламы.