Реализация критерия хи-квадрат для проверки гипотезы согласия в модели двухэтапного опроса

Кокорин Сергей Владимирович, 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
2008г.

Постановка задачи

Есть параметрическая дискретная модель P_θ данных с параметрами $\theta = (\theta_1, \ldots, \theta_r)$, задаваемая вероятностями состояний как $\{p_k(\theta)\}$, $k = 1, \ldots, s$.

Задача: проверить по критерию χ^2 согласие данных с моделью.

Проблемы:

- Получение о.м.п. (или асимптотически эквивалентных им) для параметров модели θ.
- Как достичь выполнения условия $np_k(\theta) \geq 5$ при фиксированном n.

Схема решения

1. Развитие библиотеки на C++, с помощью которой можно численно получать оценки асимптотически эквивалентные о.м.п. для дискретной модели.
2. Для конкретной модели, модели двухэтапного опроса, рассматриваются разные способы объединения состояний и реализуется способ построения критерия.

Одношаговые оценки

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

$$l(\theta, X) = \sum n_k \ln p_k(\theta).$$

Предполагаем: выполнены условия регулярности оценок максимального правдоподобия.
Информационная матрица с элементами

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

Одношаговые оценки:

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

Построение оценок а.э.о.м.п.

Одношаговые оценки:

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

Хотим: $\mathcal{L}\left(\sqrt{n}(\hat{\theta} - \theta^0)\right) \Rightarrow \mathcal{N}(\mathbf{0}, I^{-1}(\theta^0))$.

Достаточное условие на начальные оценки:
$\hat{\theta}$ — а.э.о.м.п. оценка, если $\bar{\theta}$ такая, что

$$\mathcal{L}\left(\sqrt{n}(\bar{\theta} - \theta^0)\right) \Rightarrow \mathcal{N}(0, \Sigma),$$

где Σ — невырожденная ковариационная матрица.

Модель двухэтапного опроса

Параметры

- p_0 — доля рекламируемого продукта на рынке,
- p_i, $i = 1, \ldots, m-2$ — доли остальных продуктов, $(p_{m-1} = 1 - \sum_{i=0}^{m-2} p_i)$,
- p_{ch} — вероятность хаотичного выбора после рекламы,
- p_{adv} — вероятность эффективности рекламы.

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

В качестве параметрического множества будем брать $\{\theta = (p_0, \ldots, p_{m-2}, p_{adv}, p_{ch}) \in \Theta \subset \mathbb{R}^r\}$, где $r = m + 1$. При этом будем предполагать, что истинное значение параметра θ^0 с некоторой окрестностью лежит в Θ.

Первый способ группировки

Для проверки гипотезы согласия модели используем метод χ^2.

$$X^2 = \sum_{i,j=0}^{m-1} \frac{(n_{ij} - np_{ij}(\hat{\theta}))^2}{np_{ij}(\hat{\theta})}, \quad \mathcal{L}\left(X^2\right) \Rightarrow \chi^2(m^2 - (m+1) - 1).$$

Группировка по состояниям:
Пусть $\cup_{i=0}^{t-1} I_t = \{(0,0), \ldots, (m-1, m-1)\}$, тогда новые состояния

$$q_k(\theta) = \sum_{(i,j)\in I_k} p_{ij}(\theta), \quad k = 0, \ldots, t-1.$$

- Набор параметров не меняется.
- Число состояний уменьшается: $t \leq m^2$.
- В качестве начальных оценок $\bar{\theta}$ группированной модели берутся начальные оценки исходной модели.
- X^2 приближённо распределена как $\chi^2(t - (m+1) - 1)$.

Второй способ группировки

Группировка по продуктам:
m продуктов $\longrightarrow$ w составных продуктов.

Свойство модели:
Пусть $\cup_{u=0}^{w-1} J_u = \{0, \ldots, m-1\}$, $J_u \cap J_v = \emptyset, u \neq v$, тогда

$$p'_{uv}(\theta) = \sum_{i \in J_u, j \in J_v} p_{ij}(\theta), \quad u, v = 0, \ldots, w-1$$

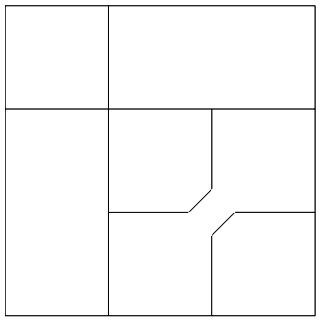
являются вероятностями для модели двухэтапного опроса с параметрами
$\theta' = (\{\sum_{i \in J_k} p_i : k = 0, \ldots, t-1\}, p_{ch}, p_{adv})$

- Количество параметров уменьшается с $m+1$ до $w+1$.
- Число состояний уменьшается с m^2 до w^2.
- X^2 приближённо распределена как $\chi^2(w^2 - (w+1) - 1)$, $w \leq m$. Следовательно, w должно быть не меньше 3.

Максимальная группировка

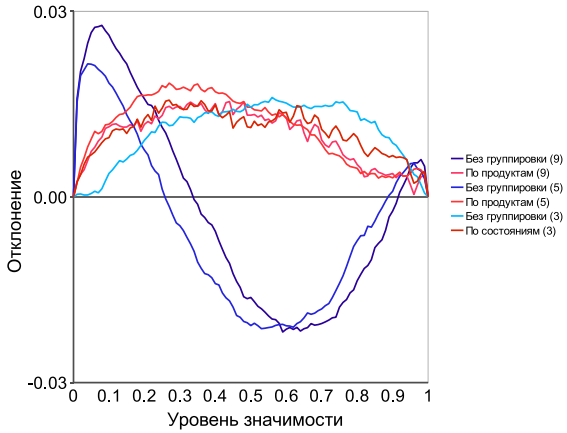
Максимальная группировка:

1. Группировка по продуктам — 3 "новых" продукта по схеме $(0, 1, 2, 1, 2, \ldots)$
2. Группировка по состояниям.

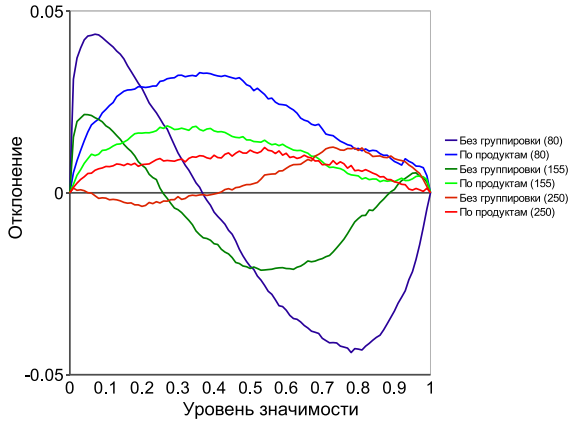


Последовательное применение группировки по продуктам и группировки по состояниям позволяет свести модель к варианту:
3 продукта, 6 состояний, 4 параметра с распределением статистики критерия $\chi^2(1)$.

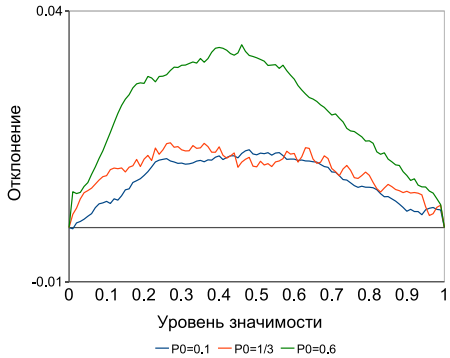
Отклонение от заданного уровня значимости:



155 индивидов, разное число продуктов



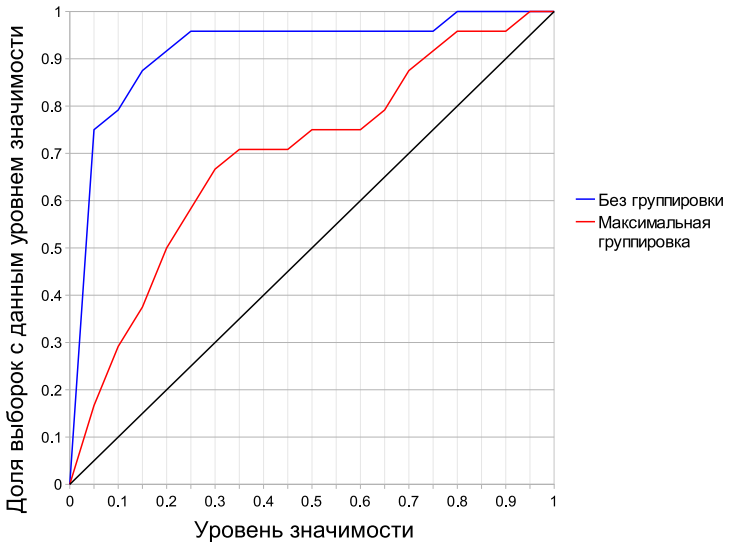
5 продуктов, разное число индивидов



155 индивидов, 3 продукта,

разные доли рекламируемого товара

Обработка реальных данных



Согласие реальных данных с моделью

26 таблиц.
$p_0 \in [0.01; 0.7]$,
$p_{ch} \in [0.15; 0.4]$,
$p_{adv} \in [0; 0.1]$.
Количество индивидов от 170 до 330.
Количество продуктов от 7 до 18.

Схема реализованной библиотеки