Welcome back my friends to the show that never ends

Критерий отношения правдоподобий. Проверка линейных гипотез

Чертоги разума

Пусть $X_1, \ldots, X_n \sim \mathsf{P}_\theta$, где θ берётся из k-мерной поверхности Θ. Поставим на проверку гипотезу $H_0\colon \theta \in \Theta_0$, где $\Theta_0 \subset \Theta$ — подповерхность размерности $l < k$. В качестве статистики критерия можно взять обобщение статистики из р.н.м. критерия Неймана-Пирсона:

$$LR(\mathbf{X}) = \frac{\sup_{\theta \in \Theta} \rho_\theta(\mathbf{X})}{\sup_{\theta \in \Theta_0} \rho_\theta(\mathbf{X})}.$$

Сам *критерий отношения правдоподобий (КОП)* имеет вид $\{\mathbf{x}\colon LR(\mathbf{x}) > c\}$. Осталось лишь подобрать c так, чтобы критерий имел нужный уровень значимости. Теорема ниже позволяет контролировать его в асимптотическом смысле:

Теорема (Уилкс). В некоторых условиях регулярности при верности нулевой гипотезы:

$$2 \ln LR(\mathbf{X}) \xrightarrow{d} \chi^2_{k-l}.$$

То есть критерий можно взять в виде $\{\mathbf{x}\colon 2\ln LR(\mathbf{x}) > \chi^2_{k-l, 1-\alpha}\}$.

Пример. $X_i \sim \mathrm{Cat}(p_1, \ldots, p_k)$, $H_0\colon \mathbf{p} \in \Theta_0$, $\dim \Theta_0 = l$. Тогда

$$2 \ln LR(\mathbf{X}) = \inf_{\mathbf{p} \in \Theta_0} 2 \sum_{i=1}^{k} \nu_i \ln \frac{\nu_i}{np_i} \xrightarrow{d} \chi^2_{k-1-l}, \quad \nu_i = \sum_{j=1}^{n} I(X_j = i).$$

Пример (линейная гипотеза). Рассмотрим гауссовскую линейную регрессию $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, $\mathbf{X} \in \mathbb{R}^n$, $\boldsymbol{\theta} \in \mathbb{R}^k$. Поставим на проверку линейную гипотезу $H_0\colon T\boldsymbol{\theta} = \boldsymbol{\tau}$, где $T \in \mathbb{R}^{m \times k}$ $(m < k)$ — полного ранга. То есть проверяется принадлежность θ какому-то $(k-m)$-мерному аффинному подпространству $\mathcal{L}_0$. Пусть $\widehat{\theta}$ — МНК-оценка, а $\widetilde{\theta}$ — МНК-оценка при верности H_0 (a.k.a. такой вектор, что $Z\widetilde{\theta} = \mathrm{proj}_{\mathcal{L}_0}\mathbf{X}$). Тогда

$$2 \ln LR(\mathbf{X}) = n \cdot \ln \frac{RSS_0}{RSS}, \quad \text{где } \ RSS = \|\mathbf{X} - Z\widehat{\theta}\|^2, \ RSS_0 = \|\mathbf{X} - Z\widetilde{\theta}\|^2.$$

КОП можно сделать точным, если заметить, что

$$RSS \sim \chi^2_{n-k} \perp\!\!\!\perp RSS_0 - RSS = \|Z\widehat{\theta} - Z\widetilde{\theta}\|^2 \sim \chi^2_m$$

как квадраты длин проекций на ортогональные подпространства. Их отношение с точностью до константы имеет *распределение Фишера* с параметрами m и $n-k$:

$$F(\mathbf{X}) = \frac{(RSS_0 - RSS)/m}{RSS/(n-k)} = \frac{\chi^2_m/m}{\chi^2_{n-k}/(n-k)} \overset{\text{def}}{\sim} F_{m, n-k},$$

откуда КОП будет выглядеть так: $\{\mathbf{x}\colon F(\mathbf{x}) > f_{m, n-k, 1-\alpha}\}$.

مسائل

1. Пусть $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \dots, n$, независимы в совокупности. Покажите, что статистика КОП для проверки $H_0\colon \mu = \mu_0$ совпадает со статистикой критерия, построенного по точному ДИ для μ.

2. Пусть $\mathbf{X} = (X_1, \dots, X_n)$ — выборка из распределения $\mathrm{U}[0; \theta]$. Чему равна статистика критерия отношения правдоподобий для проверки гипотезы $H_0\colon \theta = 1$ против общей альтернативы? Каково её предельное распределение, а каково должно быть по теореме Уилкса?

3. Среди 2020 семей, имеющих двух детей, у 527 семей по 2 мальчиков, а у 476 — по 2 девочек. Можно ли на уровне значимости 0.05 считать, что полы старшего и младшего ребёнка независимы и одинаково распределены? Сформулируйте задачу на языке КОП и воспользуйтесь теоремой Уилкса для построения критерий.

4. Команда курса практикума по статистике заметила, что в первые $2\frac{6}{7}$ недели из трёх до дедлайна в чате почти нет вопросов. По первым трём дедлайнам (получилось $20 \times 3 = 60$ дней) они записали статистику того, сколько вопросов поступило в чат в тот или иной день (для простоты считаем эти величины независимыми):

Кол-во вопросов	0	1	2	3+
Число дней	17	19	17	7

Так как каждый студент, коих много, задаёт вопрос с какой-то вероятностью, коя мала, логично выдвинуть гипотезу, что распределение числа вопросов подчиняется Пуассону (в силу одноимённой предельной теоремы). Помогите проверить сие предположение на уровне значимости 0.05 с помощью КОП. Чему будет равно число степений свобод к предельного распределения?

Замечание. Возможно, вам придётся написать пару троек строчек на Python.

5. В модели линейной регрессии $X_i = \theta_0 + \mathbf{z}_i\boldsymbol{\theta} + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\theta_0, \boldsymbol{\theta} \in \mathbb{R}^k, \sigma^2$ — неизвестные параметры. Постройте точный критерий для проверки адекватности модели $H_0\colon \boldsymbol{\theta} = \mathbf{0}$ вида $\{R^2 > c_\alpha\}$, где R^2 — *коэффициент детерминации*:

$$R^2 = 1 - \frac{\sum_{i=1}^n (X_i - \widehat{\theta}_0 - \mathbf{z}_i\widehat{\boldsymbol{\theta}})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

6. Пусть имеется k независимых выборок $\mathbf{X}_1, \dots, \mathbf{X}_k$, причём i-ая из них

$$\mathbf{X}_i = (X_{i1}, \dots, X_{in_i})$$

состоит из n_i наблюдений, которые распределены как $\mathcal{N}(\mu_i, \sigma^2)$ (дисперсии считаются равными, хоть и неизвестными). Сведя задачу к линейной регрессии, постройте точный критерий для проверки гипотезы $H_0\colon \mu_1 = \mu_2 = \ldots = \mu_k$ против общей альтернативы.