Исследование возможностей дискриминации регрессионных моделей методами перестановок

Одинцова Ольга Александровна, гр. 522

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

Научный руководитель: д.ф.-м.н., проф. Мелас В.Б.
Рецензент: к.ф.-м.н., доц. Шпилев П.В.

Санкт-Петербург
2014г.

Задача дискриминации регрессионных моделей

Результаты наблюдений описываются соотношениями

$$Y_{ij}(t_k) = f_i(t_k) + \epsilon_{ij}(t_k), \quad i = 1, 2, \quad j = 1, \ldots, n_i, \quad k = 1, \ldots, K,$$

где $Y_{ij}(t)$ — j - ая кривая в i - ой группе, и выполняются следующие предположения:

- $f_i : [a, b] \to \mathbb{R}^1, \quad i = 1, 2$ — неизвестные вещественные функции,
- $\epsilon_{ij}(t_k), \quad i = 1, 2, \quad j = 1, \ldots, n_i$ — независимые одинаково распределенные случайные величины с нулевым средним,
- $\mathsf{M}(Y_{ij}(t)) = f_i(t)$.

Задача: проверка гипотезы

$$H_0 : f_1 = f_2$$

при альтернативе

$$H_1 : f_1 \neq f_2.$$

Перестановочные критерии

- Для $t = t_1, \dots, t_K$ определим векторы

$$Z(t, \pi_0) = \{Y_{11}(t), \dots, Y_{1n}(t), Y_{21}(t), \dots, Y_{2n}(t)\},$$

$$Z(t, \pi_k) = \{\tilde{Y}_{11}(t), \dots, \tilde{Y}_{1n}(t), \tilde{Y}_{21}(t), \dots, \tilde{Y}_{2n}(t)\},$$

$$\tilde{Y}_{1i_l} = Y_{2j_l}, \quad \tilde{Y}_{2i_l} = Y_{1j_l}, \quad l = 1, \dots, k,$$

$$\tilde{Y}_{1j} = Y_{1j}, \quad j \neq i_1, \dots, i_k,$$

$$\tilde{Y}_{2j} = Y_{2j}, \quad j \neq j_1, \dots, j_k,$$

  где $\pi_k = \pi_k(s), s = 1, \dots, (C_n^k)^2$ — различные способы замены k кривых из первой группы на k кривых из второй группы.
- Определим критерий $K = K(Z)$.
- Для $Z = Z(\pi_0)$ и $Z = Z(\pi), \pi = \pi_k(s), s = 1, \dots, (C_n^k)^2$, $k = 1, \dots, n-1$ найдем значение критерия $K = K(Z)$.
- Вычислим p - value:

$$p = \frac{1}{\sum_{k=1}^{n-1} (C_n^k)^2} \sum_{\pi} \mathcal{I}(K(Z(\pi)) > K(Z(\pi_0)),$$

  где $\mathcal{I}()$ — индикаторная функция.

Критерии, не использующие попарные сравнения

Выборочная средняя кривая и выборочная медиана для i-ой группы:

$$\bar{f}_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}(t), \quad \widetilde{f}_i(t) = \mathrm{med}_{j\in n_i} Y_{ij}(t).$$

Критерии:

$$\overline{S}_1(Z) = \frac{1}{K}\sum_{k=1}^{K}\left|\overline{f}_1(t_k) - \overline{f}_2(t_k)\right|, \quad \widetilde{S}_1(Z) = \frac{1}{K}\sum_{k=1}^{K}\left|\widetilde{f}_1(t_k) - \widetilde{f}_2(t_k)\right|,$$

$$S_2(Z) = \frac{1}{K}\sum_{k=1}^{K}\left|\bar{f}_1(t_k) - \bar{f}_2(t_k)\right|^2,$$

$$S_\infty(Z) = \max_{k=1,\ldots,K}\left|\widetilde{f}_1(t_k) - \widetilde{f}_2(t_k)\right|,$$

$$S_A(Z) = \frac{1}{K}\sum_{k=1}^{K}\left|\bar{f}_1(t_k) - \bar{f}_2(t_k)\right|.$$

Критерии, использующие попарные сравнения

Вид критерия:

$$\overline{T}_1(Z) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \delta(Y_{1i}, Y_{2j}),$$

в качестве меры расстояния $\delta(Y_{1i}, Y_{2j})$ рассмотрим

$$\delta_1(Y_{1i}, Y_{2j}) = \frac{1}{K} \sum_{k=1}^{K} \left| Y_{1i}(t_k) - Y_{2j}(t_k) \right|,$$

$$\delta_2(Y_{1i}, Y_{2j}) = \frac{1}{K} \sum_{k=1}^{K} \left| Y_{1i}(t_k) - Y_{2j}(t_k) \right|^2,$$

$$\delta_\infty(f_1, f_2) = \max_{k=1,\dots,K} \left| Y_{1i}(t_k) - Y_{2j}(t_k) \right|,$$

$$\delta_A(Y_{1i}, Y_{2j}) = \frac{1}{K} \left| \sum_{k=1}^{K} (Y_{1i}(t_k) - Y_{2j}(t_k)) \right|.$$

Критерии, использующие попарные сравнения

В качестве суммарной статистики рассмотрим:

- среднее,
- медиану,
- 10% усеченное среднее и
- 20% усеченное среднее

Определение (l — усеченное среднее)

Пусть $\delta_{[i]}$ — i-ое значение вариационного ряда $\delta_{[1]} \leqslant \delta_{[2]} \leqslant \ldots \delta_{[n_1 n_2]}$. Целое число k — количество наблюдений, исключенных с концов вариационного ряда, $k = l n_1 n_2$, $l = 0.10, 0.20$, $k < \frac{1}{2} n_1 n_2$. l — усеченное среднее определяется формулой

$$\overline{T}_l = \frac{1}{(n_1 n_2 - 2k)} \sum_{i=k+1}^{n_1 n_2 - k} \delta_{[i]}.$$

Критерии, использующие попарные сравнения

Таблица : Обозначения для перестановочных критериев, использующих попарные сравнения.

критерий для среднего и δ_1	$\overline{T}_1$
критерий для среднего и δ_2	$\overline{T}_2$
критерий для среднего и δ_∞	$\overline{T}_\infty$
критерий для среднего и δ_A	$\overline{T}_A$
критерий для медианы и δ_1	$\widetilde{T}_1$
критерий для медианы и δ_2	$\widetilde{T}_2$
критерий для медианы и δ_∞	$\widetilde{T}_\infty$
критерий для медианы и δ_A	$\widetilde{T}_A$
критерий для 10% усеченного среднего и δ_1	$\overline{T}_{10,1}$
критерий для 10% усеченного среднего и δ_2	$\overline{T}_{10,2}$
критерий для 20% усеченного среднего и δ_1	$\overline{T}_{20,1}$
критерий для 20% усеченного среднего и δ_2	$\overline{T}_{20,2}$

Постановка задачи

Утверждения (Sirski, 2012)

- *Критерий $\overline{T}_1$ является наиболее мощным.*
- *Среднее является наилучшей суммарной статистикой.*
- *Попарные перестановочные критерии по мощности равны или превосходят аналогичные критерии, не использующие попарные сравнения.*

Задача: сравнить мощность критериев и проверить справедливость утверждений

- на моделях унимодальных и монотонных кривых из смеси Бета-распределений,
- на четырех-параметрической логистической модели.

Описание модели

Унимодальные кривые в первой и второй группе заданы в равноотстоящих точках t_k, $k = 0, \dots, 10$, на отрезке $[0, 1]$:

$$Y_{1j}(t_k) = e^{\eta_{1jk}} \mathrm{dBeta}(t_k, 2e^{a_{1j}}, 5e^{b_{1j}}),$$

$$Y_{2j}(t_k) = e^{\eta_{2jk}}[(1-p)\mathrm{dBeta}(t_k, 2\Delta_S\Delta_L e^{a_{2j}}, 5\Delta_S e^{b_{2j}}) + p\mathrm{dBeta}(t_k, 5e^{a_j^*}, 5e^{b_j^*})],$$

где

- $j = 1, \dots, 5,$
- $a_{ij}, b_{ij}, a_j^*, b_j^* \sim N(0, \sigma^2), \sigma = 0.2$ и независимы,
- η_{ijk} —мультипликативные ошибки авторегрессии.

$$p \in (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.75, 0.95),$$

$$\Delta_S \in (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0),$$

$$\Delta_L \in (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0).$$

Сравнение мер расстояния

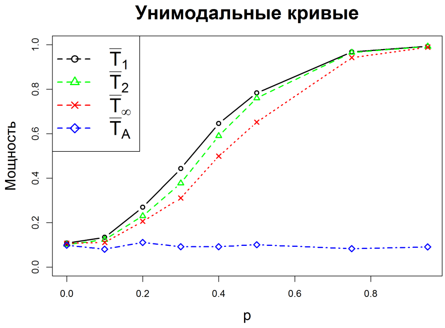

Рис. : Графики мощности для сравнения мер расстояния при среднем в качестве статистики критерия, $\alpha = 0.1$.

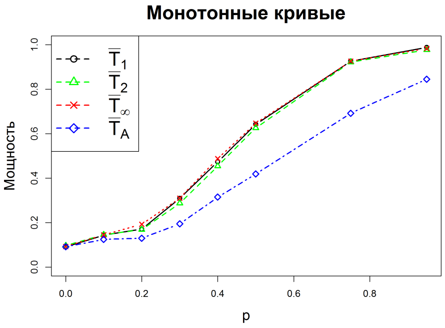

Рис. : Графики мощности для сравнения мер расстояния при среднем в качестве статистики критерия, $\alpha = 0.1$.

Критерии с использованием и без попарных сравнений

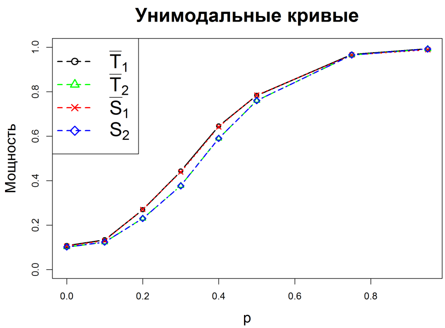

Рис. : Графики мощности для сравнения критериев, не использующих попарные сравнения, $\alpha = 0.1$.

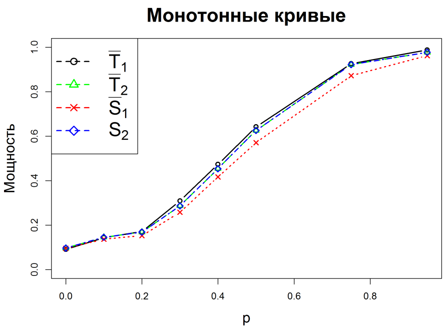

Рис. : Графики мощности для сравнения критериев, использующих и не использующих попарные сравнения, $\alpha = 0.1$.

Описание модели

Кривые в первой и второй группе заданы в равноотстоящих точках t_k, $k = 1, \ldots, 385$, на отрезке $[0, 48]$:

$$Y_{1j}(t_k) = 50 + \frac{250}{1 + \left(\frac{t_k}{\alpha_{1j}}\right)^{-2-\beta_{1j}}}, \quad j = 1, \ldots, 5$$

$$Y_{2j}(t_k) = 50 + \frac{250}{1 + \left(\frac{t_k}{\alpha_{2j} + \Delta}\right)^{-2-\beta_{2j}}}, \quad j = 1, \ldots, 5$$

где

$$\alpha_{1j} = \mathsf{U}(4, 6), \quad \alpha_{2j} = \mathsf{U}(4, 6),$$

$$\beta_{1j} = \mathsf{U}(11, 13), \quad \beta_{2j} = \mathsf{U}(11, 13),$$

$$\Delta \in (0.0, 0.37, 0.75, 1.12, 1.50, 1.87, 2.25, 2.62, 3.00).$$

Сравнение мер расстояния

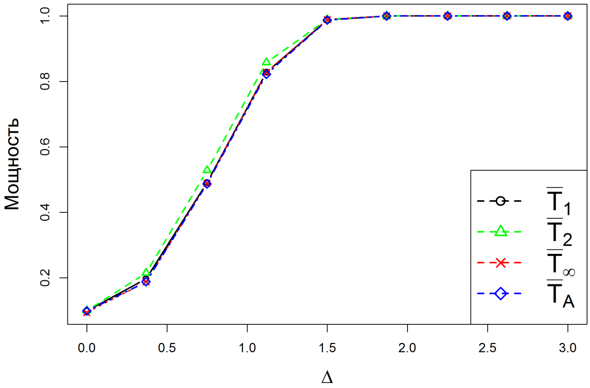

Рис. : Графики мощности для сравнения мер расстояния при среднем в качестве статистики критерия, $\alpha = 0.1$.

Сравнение суммарных статистик

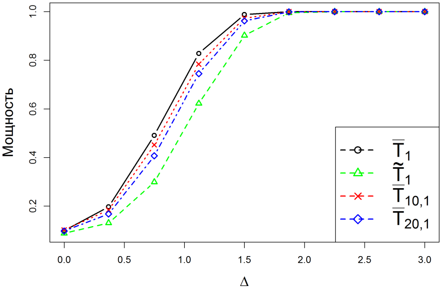

Рис. : Графики мощности для сравнения суммарных статистик при δ_1 в качестве меры расстояния, $\alpha = 0.1$.

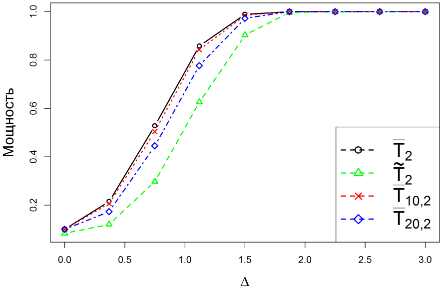

Рис. : Графики мощности для сравнения суммарных статистик при δ_2 в качестве меры расстояния, $\alpha = 0.1$.

Критерии с использованием и без попарных сравнений

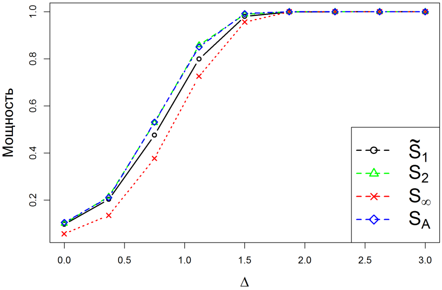

Рис. : Графики мощности критериев, не использующих попарные сравнения, $\alpha = 0.1$.

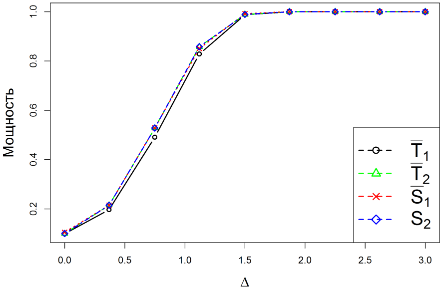

Рис. : Графики мощности критериев с использованием и без попарных сравнений, $\alpha = 0.1$.

Итоги

- Критерии и моделирование были реализованы в программной среде R.
- Полученные численные результаты позволяют сделать выводы о том, что
  - утверждения [Sirski, 2012] в целом подтверждаются на моделях из смеси Бета-распределений,
  - на четырех-параметрической логистической модели результаты отличаются от утверждений, но разница не велика,
  - рассмотренные дискретные критерии применимы для дискриминации регрессионных моделей.

Спасибо за внимание!