Дельта-метод и его некоторые применения в статистике

Вирко Елизавета Петровна, гр. 422

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

Научный руководитель: к.ф.-м.н., доцент В. В. Некруткин
Рецензент: исследователь Е. А. Советкин

Санкт-Петербург
2019г.

Одномерная модель линейной регрессии с фиксированными регрессорами

Модель:
при $i = 1, \ldots, n \quad y_i = ax_i + b + \varepsilon_i$,

- x_i — известные постоянные числа (“регрессоры”),
- ε_i — независимые одинаково распределенные случайные величины, $\mathrm{E}\,\varepsilon_i = 0, \mathrm{D}\,\varepsilon_i = \sigma^2$,
- y_i — результаты наблюдений.

Применение МНК приводит к оценкам

$$\widehat{a}_n = \frac{\mathrm{cov}_n}{\overline{s}^2_{xn}}, \quad \widehat{b}_n = \overline{y}_n - \widehat{a}_n \overline{x}_n, \quad \widehat{\sigma}^2_n = \overline{s}^2_{yn} - \widehat{a}^2_n \left(\overline{s}^2_{xn}\right)^2.$$

Некоторые свойства:

- $\widehat{a}_n, \widehat{b}_n, (n-2)\widehat{\sigma}^2_n/n$ — несмещенные ([Ивченко, Медведев, 2010]).
- Если $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, то $\widehat{\sigma}^2_n$ не зависит от $(\widehat{a}_n, \widehat{b}_n)^{\mathrm{T}}$ и известны распределения $(\widehat{a}_n, \widehat{b}_n)^{\mathrm{T}}$, $\widehat{\sigma}^2_n$.

Одномерная модель линейной регрессии с условными математическими ожиданиями

Обычно имеем дело с повторной независимой выборкой.

Модель: пусть x, y — случайные величины с конечными вторыми моментами, $\mathrm{D}x = \sigma_x^2 > 0$, такие, что $\mathrm{E}(y \,|\, x) = ax + b$. Тогда

$$a = \frac{\mathrm{cov}(x, y)}{\sigma_x^2}, \quad b = \mathrm{E}\, y - a\, \mathrm{E}\, x, \quad \sigma^2 = \sigma_y^2 - \mathrm{cov}^2(x, y)/\sigma_x^2.$$

$(x_1, y_1)^{\mathrm{T}}, \ldots, (x_n, y_n)^{\mathrm{T}}$ — повторная независимая выборка из $(x, y)^{\mathrm{T}}$. Оценки параметров формально совпадают с полученными для модели с фиксированными регрессорами.

Некоторые свойства:

- $\widehat{a}_n, \widehat{b}_n$ — несмещенные ([Демиденко, 1981]).
- Если $\mathrm{E}(\varepsilon^2 \,|\, x) = \mathrm{const}$, то $\mathrm{E}\, \widehat{\sigma}_n^2 = n\sigma^2/(n-2)$, известно асимптотическое распределение $(\widehat{a}_n, \widehat{b}_n)^{\mathrm{T}}$.

Проверка гипотезы $b = 0$. Стандартный критерий

На предыдущих двух моделях регрессии основан **стандартный критерий**.

Нулевая гипотеза: $H_0 : b = 0$. Статистика:

$$\tau_{\mathsf{std}} = \sqrt{n}\,\widehat{b}_n \,/\, \widehat{\sigma}_{\mathsf{std}}, \quad \text{где } \widehat{\sigma}_{\mathsf{std}} = \frac{n}{n-2}\,\frac{\widehat{\sigma}_n^2}{\overline{s}_{xn}^2}\,\overline{(x^2)}_n.$$

Распределение статистики при $b = 0$:

- x_i фиксированы, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. Тогда $\mathcal{L}(\tau_{\mathsf{std}}) = t(n-2)$ ([Rao et al., 2007], [Ивченко, Медведев, 2010]).
- $\mathrm{E}(y \,|\, x) = ax + b$ и $\mathrm{E}((y - ax - b)^2 \,|\, x) = \mathrm{const}$. Тогда $\mathcal{L}(\tau_{\mathsf{std}}) \Rightarrow \mathrm{N}(0, 1)$ [Демиденко, 1981].
- $(x, y)^{\mathrm{T}}$ имеет невырожденное гауссовское распределение. Тогда $\mathcal{L}(\tau_{\mathsf{std}}) = t(n-2)$.

Согласно этим распределениям строится критерий, называемый **стандартным**. Именно он реализован в пакетах STATISTICA, SPSS, STATGRAPHICS и stats для языка R.

Одномерная модель линейной регресии с линейной аппроксимацией в $\mathbb{L}^2(d\mathrm{P})$. Постановка задачи

Проблемы:

- При каких еще условиях стандартный критерий применим (асимптотически точен)?
- Что делать, если стандартный критерий не применим?

Чтобы ответить на эти вопросы — более общая модель регрессии — линейная аппроксимация в $\mathbb{L}^2(d\mathrm{P})$. Снова повторная независимая выборка. Параметры и их МНК оценки те же, что в модели с УМО. Но без условия $\mathrm{E}(y \,|\, x) = ax + b$.

Задача:

- Найти условия применимости стандартного критерия в случае линейной аппроксимации в $\mathbb{L}^2(d\mathrm{P})$.
- Построить критерии, применимые в более общей ситуации.

Обозначим $x^* = (x - \mathrm{E}x)/\sigma_x$, $\varepsilon = y - ax - b$.

Предложение

Пусть $(x, y)^{\mathrm{T}}$ *обладает конечными четвертыми моментами, распределение* x *непрерывно. Тогда* $\mathcal{L}\Big(\sqrt{n}\big((\widehat{a}_n, \widehat{b}_n, \widehat{\sigma}^2_n)^{\mathrm{T}} - (a, b, \sigma^2)^{\mathrm{T}}\big)\Big) \Rightarrow \mathrm{N}(\mathbf{0}, \Sigma)$, *где* Σ *имеет вид*

$$\frac{1}{\sigma_x^2}\begin{pmatrix} \mathrm{D}(\varepsilon x^*) & \mathrm{E}(\varepsilon^2 x^*)\sigma_x - \mathrm{E}x\mathrm{D}(\varepsilon x^*) & \mathrm{E}(\varepsilon^3 x^*)\sigma_x \\ \mathrm{E}(\varepsilon^2 x^*)\sigma_x - \mathrm{E}x\mathrm{D}(\varepsilon x^*) & \mathrm{D}(\varepsilon(\sigma_x - x^*\mathrm{E}x)) & \sigma_x^2\mathrm{E}\varepsilon^3 - \mathrm{E}(\varepsilon^3 x^*)\sigma_x \\ \mathrm{E}(\varepsilon^3 x^*)\sigma_x & \sigma_x^2\mathrm{E}\varepsilon^3 - \mathrm{E}(\varepsilon^3 x^*)\sigma_x & \sigma_x^2\mathrm{D}(\varepsilon^2) \end{pmatrix}.$$

Следствие

Пусть $\mathrm{E}(\varepsilon^2 \mid x) = \mathrm{const}$. *Тогда распределение* $\sqrt{n}(\widehat{a}_n, \widehat{b}_n)^{\mathrm{T}}$ *совпадает с известным для модели с УМО:* $\mathcal{L}\Big(\sqrt{n}\big((\widehat{a}_n, \widehat{b}_n)^{\mathrm{T}} - (a, b)^{\mathrm{T}}\big)\Big) \Rightarrow \mathrm{N}(\mathbf{0}, \Sigma)$, *где*

$$\Sigma = \frac{\sigma^2}{\sigma_x^2}\begin{pmatrix} 1 & -\mathrm{E}\,x \\ -\mathrm{E}\,x & \mathrm{E}\,x^2 \end{pmatrix}.$$

Проверка гипотезы $b = 0$. Два критерия

Гипотеза $H_0 : b = 0$. Обозначим $\sigma_b^2 = \mathrm{D}\big((y - ax - b)(\mathrm{E}x^2 - x\mathrm{E}x)\big)/\sigma_x^4$.

Статистика критерия: $\tau_n = \sqrt{n}\widehat{b}_n/\widehat{\sigma}_{0n}$, где $\widehat{\sigma}_{0n}^2 \xrightarrow{\mathrm{P}_{\mathrm{H}_0}} \sigma_0^2$ и $\mathrm{P}(\widehat{\sigma}_{0n} > 0) = 1$.

Критерий

Пусть $\alpha \in (0, 1)$. Критерий отвергает H_0, если $|\tau_n| \geq C_{1-\alpha/2}$, где $C_{1-\alpha/2}$ — квантиль $\mathrm{N}(0, 1)$ уровня $1 - \alpha/2$.

Два критерия: если (общий критерий)

$$\widehat{\sigma}_{0n} = \widehat{\sigma}^2_{\mathsf{gen}} = \frac{1}{n\overline{s}^4_{xn}} \sum_{i=1}^{n} \Big((y_i - \widehat{a}_n x_i - \widehat{b}_n)(\overline{x^2_n} - x_i \overline{x}_n)\Big)^2,$$

или (модифицированный критерий)

$$\widehat{\sigma}_{0n} = \widehat{\sigma}^2_{\mathsf{mod}} = \frac{1}{n\overline{s}^4_{xn}} \sum_{i=1}^{n} \Big((y_i - \widehat{a}_n x_i)(\overline{x^2_n} - x_i \overline{x}_n)\Big)^2,$$

тогда $\mathcal{L}(\tau_n) \overset{H_0}{\Longrightarrow} \mathrm{N}(0, 1)$.

Оба критерия асимптотически точны и состоятельны против альтернативы $H_{b^*} : b = b^* \neq 0$.

О применимости стандартного критерия

Положим

$$\sigma^2_{0,\mathsf{std}} = \frac{1}{\sigma^2 \sigma_x^2 \operatorname{E} x^2} \operatorname{D} \big((y - ax - b)(\operatorname{E} x^2 - x \operatorname{E} x)\big).$$

Утверждение

Пусть x, y имеют конечные четвертые моменты, распределение x непрерывно. Стандартный критерий асимптотически точен тогда и только тогда, когда $\sigma^2_{0,\mathsf{std}} = 1$.

Следствие

1. Стандартный критерий асимптотически точен тогда и только тогда, когда $(y - ax - b)^2$ и $(x - \operatorname{E} x)^2$ некоррелированы.
2. В частности, это выполняется, если $\operatorname{E}((y - ax - b)^2 \mid x) = \mathrm{const}$ или если $\operatorname{E} x = 0$.

$\mathrm{E}(y|x) = ax + b,\ \mathrm{E}((y - ax - b)^2|x) \neq \mathrm{const},\ b = 0.$

а) $\mathrm{E}x = 0,\ \sigma^2_{0,\mathsf{std}} = 1$: стандартный критерий асимптотически точен,

б) $\mathrm{E}x = 2,\ \sigma^2_{0,\mathsf{std}} \approx 0.85^2$: стандартный критерий асимптотически консервативен.

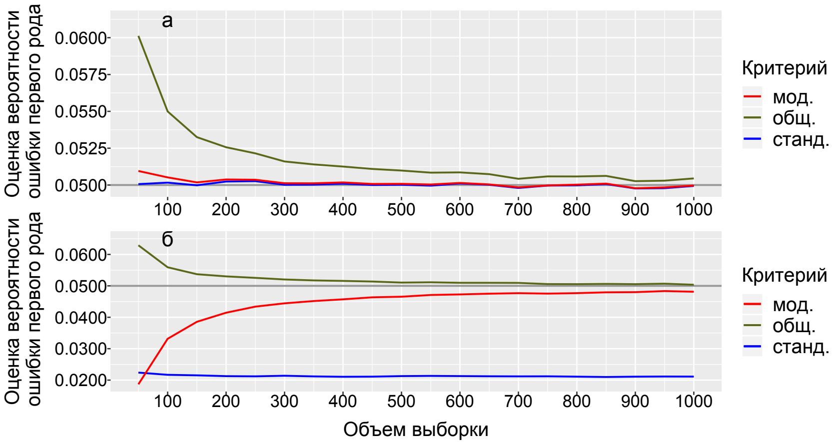

Рис.: Оценки вероятностей ошибок первого рода для равномерного распределения в эллипсе, $n = 50\,(50)\,1000$, $\alpha = 0.05$.

Поведение общего и модифицированного критериев при сдвигах

Пусть $\mathrm{E}\, x = \mathrm{E}\, y = 0$, тогда $b = 0$.
Для $c \neq 0$ положим $x' = x + c$, $y' = y + ac$. Тогда $b' = 0$.
Зафиксируем n и $(x_1, y_1)^{\mathrm{T}}, \ldots, (x_n, y_n)^{\mathrm{T}}$, и рассмотрим статистики критериев, построенные по $(x_1', y_1')^{\mathrm{T}}, \ldots, (x_n', y_n')^{\mathrm{T}}$ как функции от $c = \mathrm{E} x'$.

Лемма

1. Статистика общего критерия при $|c| \to \infty$ имеет вид

$$-\sqrt{n}\,\mathrm{sgn}(c)(\widehat{a}_n - a)/\widehat{\theta} + \mathrm{O}(1/|c|),$$

где

$$\widehat{\theta} = \frac{1}{\sqrt{n}\overline{s}_{xn}^2} \left(\sum_{i=1}^{n} (y_i - \widehat{a}_n x_i - \widehat{b}_n)^2 (\overline{x}_n - x_i)^2 \right)^{1/2}.$$

2. Статистика модифицированного критерия при $|c| \to \infty$ имеет вид

$$-\sqrt{n}\overline{s}_{xn}/c \;+\; \mathrm{O}(1/c^2).$$

Ошибки первого рода. Влияние сдвига

$\mathrm{E}(y|x) = ax + b$, $\mathrm{E}((y - ax - b)^2|x) \neq \mathrm{const}$, $b = 0$.
$\mathcal{L}(\tau_{\mathsf{std}}) \Rightarrow \mathrm{N}(0, \Delta^2)$, $\quad \Delta^2 = 2/3 + \mathrm{O}(1/c^2), |c| \to \infty$.

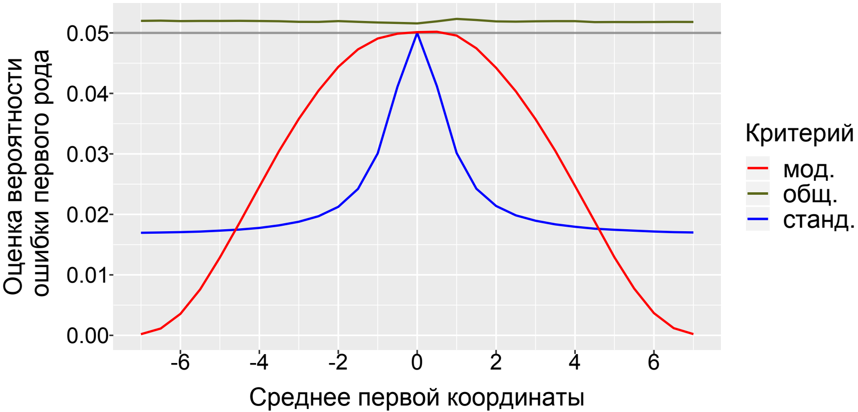

Рис.: Оценки вероятностей ошибок первого рода для равномерного распределения в эллипсе с различными сдвигами $c = \mathrm{E}\,x = -7\,(0.5)\,7$ при фиксированном объеме выборки $n = 300$, $\alpha = 0.05$.

Двумерная линейная регрессия. Три модели

- $y_i = a_1 x_{i1} + a_2 x_{i2} + b + \varepsilon_i$,
  x_{i1}, x_{i2} фиксированы, ε_i — i.i.d., $\mathrm{E}\,\varepsilon_i = 0$, $\mathrm{D}\,\varepsilon_i = \sigma^2$.
  $(\widehat{a}_{1n}, \widehat{a}_{2n}, \widehat{b}_n, \widehat{\sigma}^2_n)^{\mathrm{T}}$ — оценки МНК.

- x_1, x_2, y — случайные величины, $\mathrm{E}(y\,|\,x_1, x_2) = a_1 x_1 + a_2 x_2 + b$,
  $(x_{11}, x_{12}, y_1)^{\mathrm{T}}, \ldots, (x_{n1}, x_{n2}, y_n)^{\mathrm{T}}$ — повторная независимая выборка, оценки МНК формально те же.

Свойства первых двух моделей переносятся с одномерного случая.

- Снова повторная независимая выборка, отказ от
  $\mathrm{E}(y\,|\,x_1, x_2) = a_1 x_1 + a_2 x_2 + b$, параметры и оценки те же.

Проверка гипотезы $c_1 a_1 + c_2 a_2 = 0$. Стандартный критерий

Гипотеза $H_0\colon c_1 a_1 + c_2 a_2 = 0$, где c_1, c_2 — фиксированные числа, не равные 0 одновременно.

Для обычно используемых моделей строится **стандартный критерий**.

Он точен, если

- x_{i1}, x_{i2} фиксированы, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ ([Ивченко, Медведев, 2010]),
- $(x_1, x_2, y)^{\mathrm{T}}$ имеет невырожденное гауссовское распределение.

Асимптотически точен, если

- $\mathrm{E}(y \,|\, x_1, x_2) = a_1 x_1 + a_2 x_2 + b$,
  $\mathrm{E}((y - a_1 x_1 - a_2 x_2 - b)^2 \,|\, x_1, x_2) = \mathrm{const}$ ([Демиденко, 1981]).

Аналогично одномерному случаю интересно провести анализ применимости стандартного критерия в случае линейной аппроксимации в $\mathbb{L}^2(d\mathrm{P})$, а также построить другие критерии.

Проверка гипотезы $c_1a_1 + c_2a_2 = 0$. Два критерия

Гипотеза $H_0\colon c_1a_1 + c_2a_2 = 0$.

В работе доказывается вариант ЦПТ для $\sqrt{n}\big((\widehat{a}_{1n}, \widehat{a}_{2n})^{\mathrm{T}} - (a_1, a_2)^{\mathrm{T}}\big)$. Он применяется к построению критериев для проверки этой гипотезы.

Статистика критерия: $\tau_n = \sqrt{n}(c_1\widehat{a}_{1n} + c_2\widehat{a}_{2n})/\widehat{\sigma}_{0n}$, где $\widehat{\sigma}^2_{0n}$ — состоятельная оценка соответствующей асимптотической дисперсии (снова ЦПТ, но при выполнении H_0).

Как и ранее, получаем две оценки — *общую* (выборочную оценку) и *модифицированную* (состоятельную, только если верна H_0), а также два соответствующий критерия.

В обоих случаях $\mathcal{L}(\tau_n) \xRightarrow{H_0} \mathrm{N}(0,1)$, и оба критерия асимптотически точны и состоятельны против альтернативы $H_c\colon c_1a_1 + c_2a_2 = c \neq 0$.

Применимость стандартного критерия

Пусть x_1, x_2, y имеют конечные четвертые моменты, распределение $(x_1, x_2)^{\mathrm{T}}$ непрерывно.

Явно выписывается необходимое и достаточное условие применимости стандартного критерия, оно состоит в равенстве единице определенной величины $\sigma^2_{0,\mathsf{std}}$.

Оно выполняется, если $\mathrm{E}((y - a_1 x_1 - a_2 x_2 - b)^2 \,|\, x_1, x_2) = \mathrm{const}$.

Гипотеза $H_0: a_1 = a_2$. Ошибки первого рода

Равномерное распределение в шаре.
$\mathrm{E}(y \,|\, x_1, x_2) = a_1 x_1 + a_2 x_2 + b$,
$\mathrm{E}((y - a_1 x_1 - a_2 x_2 - b)^2 \,|\, x_1, x_2) \neq \mathrm{const}$, $a_1 = a_2$,
$\sigma^2_{0,\text{std}} = 5/7$: стандартный критерий не применим.

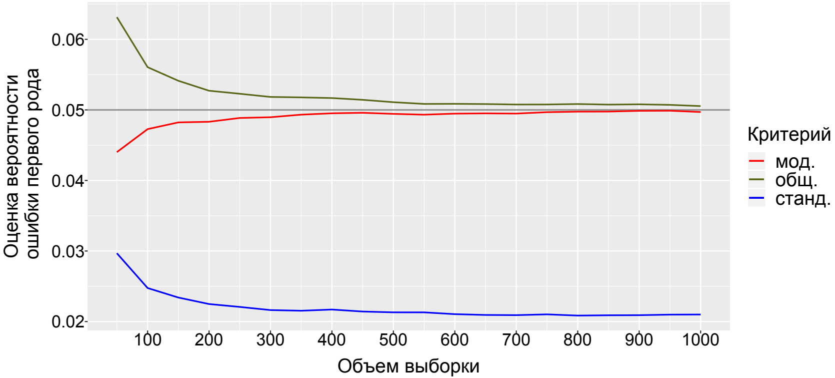

Рис.: Оценки вероятностей ошибок первого рода, $n = 50\,(50)\,1000$, $\alpha = 0.05$ для равномерного распределения в шаре.

Заключение

- Найдены асимптотические распределения оценок метода наименьших квадратов для одномерной и двумерной регрессии.
- На основе этих распределений построено по два асимптотически точных и состоятельных критерия для проверки гипотез $H_0\colon b = 0$ (в одномерном случае), $H_0\colon c_1a_1 + c_2a_2 = 0$ (в двумерном случае), изучены некоторые свойства этих критериев.
- Полученные критерии были сравнены при помощи вычислительных экспериментов между собой и с критериями, обычно применяемыми при проверке таких гипотез.
- Найдены условия применимости стандартных критериев при отказе от предположений модели с УМО.

При нахождении асимптотического распределения в общем случае использовался “дельта-метод” [Shao Jun, 1999].

Теорема

Пусть $\boldsymbol{\eta}_n$ *— последовательность* d*-мерных векторов,* $U \subset \mathbb{R}^d$ *— открытое, и* $\boldsymbol{a} \in U$*. Предположим, что отображение* $f = (f_1, \ldots, f_k)^{\mathrm{T}} : U \to \mathbb{R}^k$ *является дифференцируемым в точке* $\boldsymbol{a}$*. Обозначим* $\Delta f_i \in \mathbb{R}^d$ *— градиент функции* f_i *и положим* $\Delta_f = (\Delta f_1 : \ldots : \Delta f_k)$*.*
Пусть для некоторой последовательности $c_n \to +\infty$, $n \to \infty$ *имеет место сходимость*

$$\mathcal{L}(c_n(\boldsymbol{\eta}_n - \boldsymbol{a})) \Rightarrow \mathrm{N}(\mathbf{0}, \Sigma).$$

Пусть, к тому же, $\mathrm{P}(\boldsymbol{\eta}_n \in U) = 1$*. Тогда при* $n \to \infty$

$$\mathcal{L}\big(c_n(f(\boldsymbol{\eta}_n) - f(\boldsymbol{a}))\big) \Rightarrow \mathrm{N}(\mathbf{0}, \Sigma_f), \quad \text{где } \Sigma_f = \Delta_f^{\mathrm{T}}(\boldsymbol{a})\, \Sigma\, \Delta_f(\boldsymbol{a}).$$