Статистика

Конспект практических занятий

Мат-Мех, ПМИ, СМ–СМ

$Revision: ... (NEG: UNDER RECONSTRUCTION) $

Оглавление

[1] ANalysis Of VAriation

Часть I.

Оценки характеристик и параметров распределения

1. Выборка и эмпирическая случайная величина

Пусть $\xi \sim \mathcal{P}$ — случайная величина с распределением $\mathcal{P}$.

Определение. *Повторной независимой выборкой* объема n *(до эксперимента)* называется набор

$$\mathbf{x} = (x_1, \ldots, x_n), \quad x_i \sim \mathcal{P}\ \forall i \in 1:n,\ x_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp x_n$$

независимых в совокупности одинаково распределенных случайных величин с распределением $\mathcal{P}$.

Определение. *Повторной независимой выборкой* объема n *(после эксперимента)* называется набор реализаций, т.е. конкретных значений ξ, случайных величин x_i:

$$\mathbf{x} = (x_1, \ldots, x_n), \quad x_i \in \operatorname{supp}\xi\ \forall i \in 1:n.$$

Определение. *Эмпирической случайной величиной* $\hat{\xi}_n$ называется случайная величина с дискретным распределением

$$\hat{\xi}_n \sim \hat{\mathcal{P}}_n : \begin{pmatrix} x_1 & \ldots & x_n \\ 1/n & \ldots & 1/n \end{pmatrix}.$$

Замечание. Подходящее определение выбирается по контексту.

Если ξ имеет дискретное распределение, то выборку можно *сгруппировать*; тогда получим случайную величину $\hat{\xi}_m$ с распределением

$$\hat{\mathcal{P}}_m : \begin{pmatrix} x_1^* & \ldots & x_m^* \\ \omega_1 & \ldots & \omega_m \end{pmatrix} \quad \omega_i = \frac{\nu_i}{n},$$

где x_i^* — уникальные значения из выборки $\mathbf{x}$, а ν_i — число x_i^* в $\mathbf{x}$ (т.н. «абсолютная частота»; тогда ω_i — «относительная частота»). В противном случае, можно разбить интервал всевозможных значений выборки на m подынтервалов: $\{[e_0, e_1), \ldots, [e_{m-1}, e_m)\}$ и считать число наблюдений $\nu_i = \nu_i[e_{i-1}, e_i)$, попавших в интервал.

Следствие. *По ЗБЧ (теореме Бернулли),*

$$\omega_i \xrightarrow{\mathsf{P}} p_i = \mathsf{P}(e_{i-1} \le \xi < e_i),$$

т.е. относительная частота является хорошей оценкой вероятности на больших объемах выборки.

2. Виды признаков

Виды признаков случайной величины $\xi : \big(\Omega, \mathcal{F}, \mathsf{P}\big) \to \big(V, \mathfrak{A}\big)$ характеризуются тем, что из себя представляет множество V и что можно делать с его элементами.

Количественные признаки: $V \subset \mathbb{R}$

По типу операций:

- Аддитивные: заданы, т.е. имеют смысл в контексте данного признака, операции $+, -$
- Мультипликативные: заданы операции $\cdot, /$; признак принимает не отрицательные значения.

По типу данных:

- Непрерывные
- Дискретные

Порядковые признаки V — упорядоченное множество, определены отношения $>, =$.

Качественные признаки на V заданы отношения $=, \neq$

Пример. Цвет глаз, имена, пол.

3. Характеристики распределений и метод подстановки

Определение. *Статистика* — измеримая функция от выборки.

Обобщением статистики является понятие характеристики.

Определение. *Характеристика* — функционал от распределения:

$$T : \{\mathcal{P}\} \to V.$$

Где V — измеримое пространство, чтобы на нём можно было завести σ-алгебру.

Замечание. Чаще всего, $V = \mathbb{R}$.

Определение. Выделяют *генеральные* характеристики $T(\mathcal{P}) =: \theta$ и *выборочные* характеристики $T(\hat{\mathcal{P}}_n)$.

Определение. *Оценка* — выборочная характеристика $T(\hat{\mathcal{P}}_n) =: \hat{\theta}_n$, не зависящая от генеральной характеристики θ.

Следствие. *Выражения для вычисления генеральных и выборочных характеристик отличаются только используемыми мерами ($\mathcal{P}$ и $\hat{\mathcal{P}}_n$ соответственно).*

Определение. Пусть $\hat{\mathcal{P}}_n$ — распределение эмпирической случайной величины. Тогда *эмпирическая функция распределения* есть

$$\widehat{\mathrm{cdf}}_\xi(x) = \mathrm{cdf}_{\hat{\xi}_n}(x) = \hat{\mathcal{P}}_n((-\infty, x)) = \int_{-\infty}^{x} \mathrm{d}\hat{\mathcal{P}}_n = \sum_{x_i : x_i \le x} \frac{1}{n} = \frac{|\{x_i \in \mathbf{x} : x_i < x\}|}{n}.$$

Утверждение. Пусть $\widehat{\mathrm{cdf}}_\xi$ — эмпирическая функция распределения, cdf_ξ — функция распределения ξ. Тогда, по теореме Гливенко-Кантелли,

$$\sup_x \left| \widehat{\mathrm{cdf}}_\xi(x) - \mathrm{cdf}_\xi(x) \right| \xrightarrow{\text{a.s.}} 0.$$

Более того, если cdf_ξ непрерывна, эта сходимость имеет порядок $1/\sqrt{n}$ по теореме Колмогорова:

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{\mathrm{cdf}}_\xi(x) - \mathrm{cdf}_\xi(x) \right| \xrightarrow{\mathrm{d}} \mathcal{P}_{\text{K.S.}},$$

где $\mathcal{P}_{\text{K.S.}}$ — распределение Колмогорова-Смирнова.

Замечание. Поскольку $\widehat{\mathrm{cdf}}_\xi(x) = \omega_x$, где ω_x — частота попадания наблюдений в интервал в $(-\infty, x)$, а $\mathrm{cdf}_\xi(x) = \mathsf{P}(\xi \in (-\infty, x))$ — вероятность того же события, то можно применить теорему Бернулли (ЗБЧ):

$$\widehat{\mathrm{cdf}}_\xi(x) \xrightarrow{\mathsf{P}} \mathrm{cdf}_\xi(x).$$

Следствие. *Значит, при достаточно больших n, в качестве интересующей характеристики θ распределения $\mathcal{P}$ можем брать ее оценку $\hat{\theta}_n$ — аналогичную характеристику $\hat{\mathcal{P}}_n$.*

4. Характеристики распределений и их оценки

Определение. Генеральные и соответствующие им выборочные характеристики *k-го момента* и *k-го центрального момента*:

$$\mathsf{m}_k = \int_{\mathbb{R}} x^k \,\mathrm{d}\mathcal{P} \qquad \hat{\mathsf{m}}_k = \int_{\mathbb{R}} x^k \,\mathrm{d}\hat{\mathcal{P}}_n = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$\mathsf{m}_k^{(0)} = \int_{\mathbb{R}} (x - \mathsf{m}_1)^k \,\mathrm{d}\mathcal{P} \qquad \hat{\mathsf{m}}_k^{(0)} = \int_{\mathbb{R}} (x - \hat{\mathsf{m}}_1)^k \,\mathrm{d}\hat{\mathcal{P}}_n = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mathsf{m}}_1)^k.$$

4.1. Характеристики положения

В качестве характеристики положения выделяется 1-й момент — математическое ожидание и *выборочное среднее*:

$$\mathsf{m}_1 = \mathsf{E}\xi, \qquad \hat{\mathsf{m}}_1 =: \bar{x} = \widehat{\mathsf{E}\xi} = \mathsf{E}\hat{\xi}_n.$$

Замечание. В случае мультипликативных признаков можно посчитать среднее геометрическое; часто логарифмируют и считают среднее арифметическое.

Определение. Пусть $p \in [0,1]$ и $\mathrm{cdf} = \mathrm{cdf}_{\mathcal{P}}$. *p-квантилью* (квантилью уровня p) называется

$$\mathrm{qnt}_{\mathcal{P}}(p) =: z_p = \sup\{z : \mathrm{cdf}(z) \le p\}.$$

Квартиль есть квантиль уровня, кратного 1/4; дециль — 1/10; перцентиль — 1/100.

Замечание. sup берется для учета случая не непрерывных функций распределения.

Определение. *Медиана* есть 1/2-квантиль:

$$\mathrm{med}\,\xi = z_{1/2}.$$

Определение. Мода ($\mathrm{mode}\,\xi$) есть точка локального максимума плотности.

По методу подстановки можем получить аналогичные выборочные характеристики.

Определение. *Выборочная p-квантиль* есть такая точка $\hat{z}_p$, что она больше по значению $|\mathbf{x}|\cdot p = np$ точек из выборки:

$$\hat{z}_p = \sup\left\{z : \widehat{\mathrm{cdf}}_\xi(z) \le p\right\} = x_{(\lfloor np \rfloor + 1)}.$$

Определение. *Выборочная медиана* упорядоченной выборки $\mathbf{x} = \left(x_{(1)}, \ldots, x_{(n)}\right)$ есть

$$\hat{z}_{1/2} = \widehat{\mathrm{med}} = \begin{cases} x_{(k+1)} & n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & n = 2k \end{cases}$$

Определение. *Выборочная мода* ($\widehat{\mathrm{mode}}$) есть значение из выборки, которое чаще всего встречается.

4.2. Характеристики разброса

В качестве характеристики разброса выделяется 2-й центральный момент — дисперсия и выборочная дисперсия:

$$\mathsf{m}_2^{(0)} = \mathsf{D}\xi \qquad \hat{\mathsf{m}}_2^{(0)} =: s^2 = \widehat{\mathsf{D}\xi} = \mathsf{D}\hat{\xi}_n = \begin{cases} \mathsf{E}\left(\hat{\xi}_n - \mathsf{E}\hat{\xi}_n\right)^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ \mathsf{E}\hat{\xi}_n^2 - \left(\mathsf{E}\hat{\xi}_n\right)^2 = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2. \end{cases}$$

Замечание. Если среднее $\mathsf{E}\xi = \mu$ известно, то дополнительно вводится

$$s_\mu^2 := \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \mu)^2 \\ \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2\right) - \mu^2. \end{cases}$$

Пример (Оценка дисперсии оценки мат. ожидания)**.** Пусть строится оценка мат. ожидания $\bar{x}$. Может интересовать точность построенной оценки. Вычислим дисперсию теоретически, после чего оценим точность по выборке:

$$\mathsf{D}\bar{x} = \mathsf{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}x_i = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}\xi = \frac{\mathsf{D}\xi}{n},$$

откуда

$$\widehat{\mathsf{D}\bar{x}} = \frac{s^2}{n}.$$

Пример (Дисперсия оценки дисперсии)**.** См. по ссылке[1].

Определение (Энтропия)**.** Количество информации, необходимое для выявления объекта из n-элементного множества вычисляется по *формуле Хартли*:

$$H = \log_2 n$$

(множество это следует итеративно разбивать пополам, откуда и оценка). Пусть теперь множество не равновероятно, т.е. задано дискретное распределение

$$\mathcal{P}_\xi : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}.$$

Тогда количество информации $H(\xi)$, которую нужно получить, чтобы узнать, какой исход эксперимента осуществлен, вычисляется по формуле Шеннона и называется *энтропией*:

$$H(\xi) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}.$$

Замечание. В случае равномерного дискретного распределения, конечно, $H = H(\xi)$.

Определение. *Выборочное стандартное отклонение* есть

$$\mathrm{SD} := \sqrt{\widehat{\mathsf{D}\xi}} = s.$$

Это показатель разброса случайной величины; показатель того, насколько элементы выборки отличаются от выборочного среднего по значению.

[1] http://mathworld.wolfram.com/SampleVarianceDistribution.html

SD позволяет оценивать стандартное отклонение распределения ξ.

Пусть $\hat{\theta}_n$ — статистика. Она имеет какое-то своё распределение, стандартное отклонение которого можно также оценить.

Определение. *Стандартная ошибка* оценки есть

$$\mathrm{SE}(\hat{\theta}) := \sqrt{\widehat{\mathsf{D}\hat{\theta}}}.$$

Это показатель разброса оценки случайной величины.

Замечание. В частном случае $\theta = \mathsf{E}\xi$, $\hat{\theta} = \bar{x}$ получаем *выборочную стандартную ошибку среднего*

$$\mathrm{SE} := \mathrm{SE}(\bar{x}) = \sqrt{\widehat{\mathsf{D}\bar{x}}} = \sqrt{\frac{\widehat{\mathsf{D}\xi}}{n}} = \frac{s}{\sqrt{n}}.$$

Это, в свою очередь, показатель того, насколько выборочное среднее отличается от истинного.

Пусть $c_\gamma = \mathrm{qnt}_{\mathrm{N}(0,1)}\,\gamma$.

Пример (С мостом и машинами)**.** При возведении моста требуется, чтобы под ним могли проехать, условно, 95% машин. Чтобы эту высоту вычислить, достаточно собрать выборку высоты кузова проезжающих машин. Тогда нахождение искомой величины можно наглядно представить как выбор такой квантили гистограммы выборки, что суммирование соответствующих вероятностей даст 0.95. В предположении, что выборка из нормального распределения, с более устойчивой оценкой квантили, интервал будет иметь вид

$$\left(\bar{\mathbf{x}} \pm \mathrm{SD}\cdot c_\gamma\right).$$

SE как показатель разброса среднего использовать по смыслу нельзя.

Пример (С паромом)**.** Число машин, которое способен перевезти паром, есть Грузоподъемность/$\mathsf{E}\xi$, где ξ — вес машины. Поскольку оценка $\bar{x}$ всегда считается с погрешностью относительно истинного значения, интервал допустимого числа машин будет иметь вид

$$\frac{\text{Грузоподъемность}}{\bar{x} \pm \mathrm{SE}\cdot c_\gamma}.$$

4.3. Анализ характера разброса

Определение. *Коэффициент асимметрии Пирсона* («скошенности»[2])

$$\gamma_3 = \mathsf{A}\xi = \frac{\mathsf{m}_3^{(0)}}{\sigma^3}.$$

Замечание. Не зависит от линейных преобразований.

Замечание. Старое определение скошенности было $\frac{\mathsf{E}\xi - \mathrm{med}\,\xi}{\sigma}$.

Замечание. Типичный случай соответствует тому, что при положительном коэффициенте асимметрии ‘хвост вправо’.

Определение. *Коэффициент эксцесса* («крутизны», «kurtosis»):

$$\gamma_4 = \mathsf{K}\xi = \frac{\mathsf{m}_4^{(0)}}{\sigma^4} - 3.$$

Замечание. Величина $\mathsf{m}_4^{(0)}/\sigma^4 = 3$ соответствует стандартному нормальному распределению. Так что можно сравнивать выборку и γ_4 $\mathrm{N}(0,1)$.

Замечание. Положительный коэффициент эксцесса соответствует медленному убыванию на концах отрезка. Причём, так как распределение стандартизуется, имеется в виду убывание на хвостах, которое медленнее по порядку (!), чем убывание на хвостах у нормального распределения. Например, сравните e^{-x^2}, $e^{-x^2/10}$ и e^{-10*x}. Часто говорят об островершинности при положительном эксцессе, но это просто вторая сторона скорости убывания на хвостах.

[2] «Skewness».

4.4. Характеристики зависимости

Определение. Пусть $(\xi_1, \xi_2) \sim \mathcal{P}$ и $(x_1, y_1), \ldots, (x_n, y_n) \sim \mathcal{P}(du \times dv)$. Тогда можно записать две другие важные характеристики: *ковариацию* и *коэффициент корреляции*:

$$\operatorname{cov}(\xi_1, \xi_2) = \iint_{\mathbb{R}^2} (u - \mathsf{m}_1(\xi_1))(v - \mathsf{m}_1(\xi_2))\mathcal{P}(du \times dv) \qquad \widehat{\operatorname{cov}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\operatorname{cor}(\xi_1, \xi_2) = \frac{\operatorname{cov}(\xi_1, \xi_2)}{\sigma_{\xi_1}\sigma_{\xi_2}} \qquad \widehat{\operatorname{cor}}(\mathbf{x}, \mathbf{y}) = \frac{\widehat{\operatorname{cov}}(\mathbf{x}, \mathbf{y})}{s(\mathbf{x})s(\mathbf{y})}.$$

5. Точечная оценка параметров распределения

5.1. Метод подстановки

Метод подстановки заключается в подстановке вместо неизвестного теоретического распределения известного эмпирического распределения. Например, вас интересует некоторая характеристика $f(\xi)$, а вы в качестве оценки предлагаете $\widehat{f(\xi)} = f(\xi_n)$, где ξ_n — эмпирическая случайная величина.

5.2. Метод моментов

Пусть $\mathcal{P}(\boldsymbol{\theta}),\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^\mathsf{T}$ — параметрическая модель. Найдем оценки для параметров $\hat{\theta}_i,\ i \in \overline{1:r}$, для чего составим и решим систему уравнений:

$$\begin{cases} \mathsf{E}g_1(\xi) = \phi_1(\theta_1, \ldots, \theta_r) \\ \quad\vdots \\ \mathsf{E}g_r(\xi) = \phi_r(\theta_1, \ldots, \theta_r) \end{cases} \Longrightarrow \begin{cases} \theta_1 = f_1(\mathsf{E}g_1(\xi), \ldots, \mathsf{E}g_r(\xi)) \\ \quad\vdots \\ \theta_r = f_r(\mathsf{E}g_1(\xi), \ldots, \mathsf{E}g_r(\xi)). \end{cases}$$

Примем

$$\theta_i^* = f_i(\hat{\mathsf{E}}g_1(\xi), \ldots, \hat{\mathsf{E}}g_r(\xi)).$$

Часто, $g_i(\xi) = \xi^i$. Или, еще чаще, $g_1(\xi) = \xi$ и $g_i(\xi) = (\xi - \mathsf{E}\xi)^i$, $i > 1$, так как для таких моментов обычно известны формулы.

Замечание. Случается, что решение находится вне пространства параметров. На практике, если пространство параметров компактное, можно взять точку, ближайшую к полученной оценке. Однако это свидетельствует о том, что модель плохо соответствует данным.

Пример 5.1 ($r = 1$)**.** $\xi \sim \mathrm{U}(0, \theta)$.

- Оценка по 1-му моменту: $g(\xi) = \xi$ и

$$\mathsf{E}\xi = \int_0^\theta \frac{1}{\theta} x \,\mathrm{d}x = \frac{1}{\theta}\,\frac{x^2}{2}\bigg|_0^\theta = \frac{\theta}{2} \implies \theta = 2\mathsf{E}\xi,\ \theta^* = 2\bar{x}.$$

- Оценка по k-му моменту: $g(\xi) = \xi^k$ и

$$\mathsf{E}\xi^k = \frac{1}{\theta}\int_0^\theta x^k \,\mathrm{d}x = \frac{1}{\theta}\,\frac{x^{k+1}}{k+1}\bigg|_0^\theta = \frac{\theta^k}{k+1} \implies \theta^* = \sqrt[k]{(k+1)\frac{1}{n}\sum_{i=1}^n x_i^k}.$$

Пример 5.2 ($r = 1$)**.** Пусть $\xi \sim \mathrm{Exp}(\lambda)$. Тогда $\mathsf{E}\xi = \lambda$ и $\bar{x} = \lambda$.

5.3. Метод оценки максимального правдоподобия

Пусть $\mathcal{P}_\xi(\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^\mathsf{T}$ — параметрическая модель.

Определение. Пусть

$$\mathsf{P}(\mathbf{y} \mid \boldsymbol{\theta}) = \begin{cases} \mathsf{P}_{\boldsymbol{\theta}}(x_1 = y_1, \ldots, x_n = y_n) & \mathcal{P}_\xi(\boldsymbol{\theta}) \text{ дискретно;} \\ p_{\boldsymbol{\theta}}(\mathbf{y}) & \mathcal{P}_\xi(\boldsymbol{\theta}) \text{ абсолютно непрерывно.} \end{cases}$$

Тогда *функция правдоподобия* определяется как значение распределения выборки (плотности в непрерывном случае и вероятности значений в дискретном) с подстановкой выборки вместо аргумента:

$$\mathsf{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \mathsf{P}(\mathbf{x} \mid \boldsymbol{\theta}).$$

Пример 5.3. Пусть $\xi \sim \mathrm{N}(\mu, \sigma^2)$. По независимости x_i, $p_{\boldsymbol{\theta}}(\mathbf{x})$ распадается в произведение:

$$\mathsf{L}(\boldsymbol{\theta} \mid \mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{i=1}^n p_{\boldsymbol{\theta}}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right\}.$$

Пример 5.4. $\xi \sim \mathrm{Pois}(\lambda)$,

$$\mathsf{P}(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda} \implies \mathsf{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \prod_{i=1}^n \frac{1}{x_i!} \lambda^{x_i} e^{-\lambda} = \frac{1}{\prod_{i=1}^n x_i!} \lambda^{n\bar{x}} e^{-n\lambda}.$$

Утверждение. Пусть $\mathbf{x}$ — выборка. В качестве оценки максимального правдоподобия[1] $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ следует взять

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \log \mathsf{L}(\boldsymbol{\theta} \mid \mathbf{x}).$$

Пример. $\xi \sim \mathrm{Pois}(\lambda)$.

$$\ln \mathsf{L}(\lambda \mid \mathbf{x}) = -\sum_{i=1}^n \ln(x_i!) - n\lambda + n\bar{x} \ln \lambda \implies \frac{\partial \log \mathsf{L}(\lambda \mid \mathbf{x})}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda}$$

откуда

$$\frac{\partial \log \mathsf{L}(\lambda \mid \mathbf{x})}{\partial \lambda} = 0 \iff -n + \frac{n\bar{x}}{\lambda} = 0,\ n\bar{x} - n\lambda = 0,\ \lambda = \bar{x}.$$

Утверждение. В условиях регулярности:

1. Существует один глобальный максимум, так что

$$\left.\frac{\partial \log \mathsf{L}(\lambda \mid \mathbf{x})}{\partial \lambda}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}} = 0.$$

2. $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ обладает всеми свойствами (про определение этих свойст написано в следующих разделах):
   a) Состоятельность;
   b) Асимптотическая несмещенность;
   c) Асимптотическая нормальность;
   d) Асимптотическая эффективность.

[1] Maximum likelihood estimate (MLE).

6. Свойства оценок

6.1. Несмещенность

Определение. *Смещение*[1] есть

$$\operatorname{bias}\hat\theta_n := \mathsf{E}\hat\theta_n - \theta \quad \forall\theta\in\Theta.$$

Определение. *Среднеквадратичная ошибка*[2] есть

$$\operatorname{MSE}\hat\theta_n := \mathsf{E}(\hat\theta_n-\theta)^2.$$

Замечание. Поскольку

$$\mathsf{D}\hat\theta_n = \mathsf{D}(\hat\theta_n-\theta) = \mathsf{E}(\hat\theta_n-\theta)^2 - (\mathsf{E}(\hat\theta_n-\theta))^2,$$

то

$$\underbrace{\mathsf{E}(\hat\theta_n-\theta)^2}_{\text{MSE}} = \mathsf{D}\hat\theta_n + \underbrace{(\mathsf{E}(\hat\theta_n-\theta))^2}_{\text{bias}^2}.$$

Определение. Оценка называется *несмещенной*, если $\operatorname{bias}\hat\theta_n = 0$, т.е.

$$\mathsf{E}\hat\theta_n = \theta.$$

Предложение. *$\bar x$ — несмещенная оценка $\mathsf{E}\xi$.*

Доказательство. Пусть $\theta = \mathsf{E}\xi$, $\hat\theta_n = \mathsf{E}\hat\xi_n = \bar x$. Тогда

$$\mathsf{E}\bar x = \mathsf{E}\frac1n\sum_{i=1}^n x_i = \frac1n\sum_{i=1}^n \mathsf{E}x_i = \frac1n\sum_{i=1}^n \mathsf{E}\xi = \mathsf{E}\xi \implies \mathsf{E}\hat\theta_n = \mathsf{E}\theta,\ \operatorname{bias}\hat\theta_n = 0.$$

□

Предложение. *s^2 является только* асимптотически *несмещенной оценкой $\mathsf{D}\xi$.*

Доказательство. Это доказательство ‘в лоб’. Хорошее доказательство на основе разложения $\mathsf{E}(\zeta-a)^2$ смотрите в конспекте.

Поскольку дисперсия не зависит от сдвига, обозначим $\eta = \xi - \mathsf{E}\xi$ и $y_i = x_i - \mathsf{E}\xi$; тогда

$$\begin{aligned}
\mathsf{E}s^2 &= \mathsf{E}\widehat{\mathsf{D}\xi} = \mathsf{E}\widehat{\mathsf{D}\eta} = \mathsf{E}\left(\widehat{\mathsf{E}\eta^2} - \left(\widehat{\mathsf{E}\eta}\right)^2\right) = \mathsf{E}\left(\frac1n\sum_{i=1}^n y_i^2 - \left(\frac1n\sum_{i=1}^n y_i\right)^2\right)\\
&= \mathsf{E}\left(\frac1n\sum_{i=1}^n y_i^2 - \frac{1}{n^2}\sum_{i,j=1}^n y_iy_j\right) = \mathsf{E}\left(\frac1n\sum_{i=1}^n y_i^2 - \frac{1}{n^2}\sum_{i,j=1}^n y_i^2\right) = \frac1n\sum_{i=1}^n \mathsf{E}y_i^2 - \frac{1}{n^2}\sum_{i=1}^n \mathsf{E}y_i^2\\
&= \frac1n\sum_{i=1}^n \mathsf{E}\,(x_i-\mathsf{E}\xi)^2 - \frac{1}{n^2}\sum_{i=1}^n \mathsf{E}\,(x_i-\mathsf{E}\xi)^2 = \frac1n\sum_{i=1}^n \mathsf{D}x_i - \frac{1}{n^2}\sum_{i=1}^n \mathsf{D}x_i = \mathsf{D}\xi - \frac1n\mathsf{D}\xi\\
&= \frac{n-1}{n}\mathsf{D}\xi \xrightarrow[n\to\infty]{} \mathsf{D}\xi.
\end{aligned}$$

□

[1] Bias.
[2] Mean squared error (MSE).

Определение. *Исправленная* дисперсия:

$$\tilde{s}^2 := \frac{n}{n-1} s^2.$$

Очевидно, исправленная дисперсия — несмещенная оценка дисперсии.

6.2. Состоятельность

Определение. Оценка называется *состоятельной в среднеквадратичном смысле*, если

$$\operatorname{MSE} \hat{\theta}_n \xrightarrow[n\to\infty]{} 0.$$

Определение. Оценка называется *состоятельной*, если

$$\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta.$$

Предложение. *Если оценка несмещенная и состоятельная в среднеквадратичном смысле, то она состоятельная.*

Доказательство. В самом деле, по неравенству Чебышева,

$$\mathsf{P}(|\hat{\theta}_n - \theta| > \epsilon) = \mathsf{P}(|\hat{\theta}_n - \mathsf{E}\hat{\theta}_n| > \epsilon) \le \frac{\mathsf{D}\hat{\theta}_n}{\epsilon^2} = \frac{\operatorname{MSE} \hat{\theta}_n}{\epsilon^2} \xrightarrow[n\to\infty]{} 0.$$

□

Предложение. $\hat{\mathsf{m}}_k$ *является состоятельной оценкой* m_k.

Доказательство. Докажем для $\hat{\mathsf{m}}_1$. По определению выборки до эксперимента, $x_i \sim \mathcal{P}$. Тогда, по теореме Хинчина о ЗБЧ,

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \xrightarrow{\mathsf{P}} \mathsf{m}_1(\mathcal{P}).$$

Для k-го момента доказывается аналогично заменой $y_i := x_i^k$. □

Замечание. Для $\mathsf{m}_k^{(0)}$ доказательство не пройдет, потому что x_i и $\bar{x}$ не будут независимыми.

Предложение. $\hat{\mathsf{m}}_k^{(0)}$ *является состоятельной оценкой* $\mathsf{m}_k^{(0)}$.

Утверждение. Пусть $\xi_n \xrightarrow{\mathsf{P}} c$ и $f \in C(U_\epsilon(c))$. Тогда $f(\xi_n) \xrightarrow{\mathsf{P}} f(c)$.

Доказательство предложения. Докажем для s^2. Пусть $f : (x, y) \mapsto x - y^2$. Устроим последовательность $(\hat{\mathsf{m}}_2, \hat{\mathsf{m}}_1) \xrightarrow{\mathsf{P}} (\mathsf{m}_2, \mathsf{m}_1)$. Тогда

$$f(\hat{\mathsf{m}}_2, \hat{\mathsf{m}}_1) = \hat{\mathsf{m}}_2 - \hat{\mathsf{m}}_1^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2 = s^2 \xrightarrow{\mathsf{P}} f(\mathsf{m}_2, \mathsf{m}_1) = \mathsf{D}\xi.$$

Для $\mathsf{m}_k^{(0)}$ доказывается аналогично. □

Предложение. $\bar{x}$ — *состоятельная оценка* $\mathsf{E}\xi$.

Доказательство. Либо по (6.2) для $k = 1$, либо из того факта, что $\operatorname{bias} \bar{x} = 0$, значит

$$\operatorname{MSE} \bar{x} = \mathsf{D}\bar{x} = \frac{\mathsf{D}\xi}{n} \xrightarrow[n\to\infty]{} 0,$$

и по (6.2) получаем утверждение. □

Предложение. s^2 — *состоятельная оценка* $\mathsf{D}\xi$.

Доказательство. По (6.2) с $k = 2$. □

6.3. Асимптотическая нормальность

Определение. Оценка $\hat{\theta}_n$ называется *асимптотически нормальной* оценкой параметра θ с коэффициентом $\sigma^2(\theta)$ если

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2(\theta)).$$

Пример. $\bar{x}$ — асимптотически нормальная оценка, если $\mathsf{D}\xi < \infty,\ \mathsf{D}\xi \neq 0$:

$$\sqrt{n}\,(\bar{x} - \mathsf{E}\xi) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathsf{D}\xi).$$

Доказательство. По ЦПТ,

$$\sqrt{n}\,(\bar{x} - \mathsf{E}\xi) = \frac{\sum_{i=1}^n x_i - n\mathsf{E}\xi}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathsf{D}\xi).$$

□

Мы обсуждали, что асимптотическую нормальность можно определять и в слабом смысле — как сходимость по распределению к нормальному распределению $N(0,1)$ стандартизированной случайной величины.

6.4. Эффективность

Определение. Говорят, что оценка $\hat{\theta}^{(1)}$ *лучше* $\hat{\theta}^{(2)}$ *в среднеквадратичном смысле*, если

$$\mathrm{MSE}\,\hat{\theta}^{(1)} \leq \mathrm{MSE}\,\hat{\theta}^{(2)}.$$

Замечание. Для несмещенных оценок определение эквивалентно, конечно,

$$\mathsf{D}\hat{\theta}^{(1)} \leq \mathsf{D}\hat{\theta}^{(2)}.$$

Определение. В классе несмещенных оценок оценка называется эффективной (в средне-квадратическом), если ее дисперсия минимальна. В классе асимптотически несмещенных оценок оценка $\hat{\theta}$ называется асимптотически эффективной, если для любой другой оценки $\hat{\theta}^*$ выполнено $\mathsf{D}\hat{\theta} \leq \mathsf{D}\hat{\theta}^* \leq 1$.

6.4.1. Эффективность и неравенство Рао-Крамера

Пусть $\mathcal{P}_\xi(\boldsymbol{\theta}),\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^\mathsf{T}$ — параметрическая модель. Пусть $r = 1$.

Определение. *Информанта n-го порядка*:

$$S_n(\mathbf{x}, \theta) = \frac{\mathrm{d}^n \ln \mathsf{L}(\theta \mid \mathbf{x})}{\mathrm{d}\theta^n}.$$

Определение. *Информационное количество Фишера*:

$$I_n(\theta) := -\mathsf{E}S_2(\mathbf{x}, \theta).$$

Утверждение.

$$I_n(\theta) = \mathsf{E}S_1^2(\mathbf{x}, \theta).$$

Пример. $\xi \sim \mathrm{Pois}(\lambda)$.

$$S_1(\mathbf{x}, \theta) = -n + \frac{n\bar{x}}{\lambda}, \quad S_2(\mathbf{x}, \theta) = -\frac{n\bar{x}}{\lambda^2} \implies I_n(\lambda) = \mathsf{E}\frac{n\bar{x}}{\lambda^2} = \frac{n}{\lambda^2}\mathsf{E}\bar{x} = \frac{n}{\lambda}.$$

Замечание.

$$\ln \mathsf{L}(\theta \mid \mathbf{x}) = \sum_{i=1}^{n} \ln p_\theta(x_i) \implies S_2 = \frac{\mathrm{d}^2 \ln \mathsf{L}(\theta \mid \mathbf{x})}{\mathrm{d}\theta^2} = \sum_{i=1}^{n} (\ln p_\theta(x_i))'',$$

откуда, для повторной независимой выборки,

$$I_n(\theta) = -\sum_{i=1}^{n} \mathsf{E}(\ln p_\theta(x_i))'' = n \cdot i(\theta), \quad \text{где } i(\theta) = -\mathsf{E}(\ln p_\theta(\xi))''.$$

Определение. $C \subset \mathbb{R}$ есть *носитель* параметрического семейства распределений $\mathcal{P}(\theta)$, если

$$\xi \sim \mathcal{P}(\theta) \implies \mathsf{P}(\xi \in C) = 1, \ \forall \theta \in \Theta.$$

Определение. Условие регулярности: имеют отношение к независимости носителя распределения от параметра, а также к существованию и ограниченности производных функции логправдоподобия по параметру до определённого порядка дифференцирования.

Пример. $\mathrm{Exp}(\lambda)$ — регулярное семейство; $\mathrm{U}(0, \theta)$ — не является регулярным.

Утверждение. Для несмещенных оценок в условиях регулярности справедливо неравенство *Рао–Крамера*:

$$\mathsf{D}\hat{\theta}_n \geq \frac{1}{I_n(\theta)}.$$

Для смещенных оценок,

$$\mathsf{D}\hat{\theta}_n \geq \frac{(1 + \mathrm{bias}'(\theta))^2}{I_n(\theta)}.$$

Следствие. *Несмещенная оценка является эффективной, если:*

$$\mathsf{D}\hat{\theta}_n = \frac{1}{I_n(\theta)}.$$

Следствие. *Асимптотически несмещенная оценка является асимптотически эффективной, если:*

$$\mathsf{D}\hat{\theta}_n \cdot I_n(\theta) \to 1 \ \textit{as} \ n \to \infty.$$

Упражнение (Хорошее)**.** Показать, что $\bar{x}$ является эффективной оценкой μ в модели $\xi \sim \mathrm{N}(\mu, \sigma^2)$.

Следствие. *Пусть $\hat{\theta}_n$ — асимптотически несмещенная оценка. Тогда $\hat{\theta}_n$* — асимптотически эффективная*, если*

$$\mathsf{D}\hat{\theta}_n \cdot I_n \xrightarrow[n\to\infty]{} 1.$$

Пример. Пусть $\xi \sim \mathrm{N}(\mu, \sigma^2)$. Можно посчитать, что s^2 является только асимптотически эффективной оценкой σ^2; $\tilde{s}^2$ — просто эффективной.

Пример. Пусть $\xi \sim \mathrm{Pois}(\lambda)$. Поскольку

$$\begin{aligned} \mathsf{D}\hat{\lambda}_n &= \mathsf{D}\bar{x} = \mathsf{E}\xi/n = \lambda/n \\ I_n(\lambda) &= n/\lambda, \end{aligned}$$

то $\hat{\lambda}_n$ — эффективная оценка (как и ожидалась по свойствам $\hat{\theta}_{\mathrm{MLE}}$).

6.5. Устойчивость оценок

Так как в реальных данных часто бывают те или иные ошибки, часто жертвуют точностью для увеличения устойчивости (робастности) к выбросам. Устойчивые аналоги оценок часто строятся на основе рангов (номеров по порядку в упорядоченной выборке). Приведем пример.

Пример (Сравнение оценок мат. ожидания симметричного распределения)**.** Пусть $\mathcal{P}$ *симметрично* — в этом случае $\widehat{\operatorname{med}\xi} = \bar{x}$ и имеет смысл сравнить две этих характеристики.

$$\begin{aligned} \mathsf{D}\bar{x} &= \frac{\mathsf{D}\xi}{n} \\ \widehat{\mathsf{D}\operatorname{med}\xi} &\sim \frac{1}{4n\,\operatorname{pdf}^2_{\mathrm{N}(\mu,\sigma^2)}(\operatorname{med}\xi)} \quad \text{при } n \to \infty. \end{aligned}$$

Так, если $\xi \sim \mathrm{N}(\mu, \sigma^2)$, то

$$\operatorname{pdf}^2_{\mathrm{N}(\mu,\sigma^2)}(\operatorname{med}\xi) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{(\operatorname{med}\xi-\mu)^2}{\sigma^2}\right\} = \frac{1}{2\pi\sigma^2},$$

откуда

$$\widehat{\mathsf{D}\operatorname{med}\xi} = \frac{\pi}{2}\frac{\sigma^2}{n} > \frac{\sigma^2}{n} = \mathsf{D}\bar{x},$$

значит $\bar{x}$ эффективнее $\widehat{\operatorname{med}\xi}$.

Замечание. В то же время, $\widehat{\operatorname{med}\xi}$ более устойчива к аутлаерам, чем $\bar{x}$, и этим лучше.

Часть II.

Некоторые распределения, связанные с нормальным

1. Распределение $N(a, \sigma^2)$

Свойства хорошо известны. TODO

Здесь пока напишу только про измерение расстояния в сигмах. Будет говорить, что точка далеко от мат.ожидания, если это и более далекие значения маловероятны.

Формально, пусть $\xi \sim N(a, \sigma^2)$. Рассмотрим $\mathsf{P}(|\xi - a| > k\sigma)$. Эта вероятность не зависит от σ и равна $2(1 - \Phi(k))$, где $\Phi(x)$ — функция стандартного нормального распределения.

Значения $\mathsf{P}(|\xi - a| > k\sigma)$:

k	Вероятность
1	0.317
1.64	0.101
1.96	0.050
2	0.046
3	2.70E-03
6	1.97E-09

2. Распределение $\chi^2(m)$

Определение (Распределение $\chi^2(m)$)**.** η имеет распределение χ^2 с m степенями свободны:

$$\eta \sim \chi^2(m) \iff \eta = \sum_{i=1}^{m} \zeta_i^2, \quad \zeta_i \sim \mathrm{N}(0,1),\ \zeta_i \text{ независимы.}$$

Свойства[1] $\chi^2(m)$

$$\begin{aligned} \mathsf{E}\eta &= \sum_{i=1}^{m} \mathsf{E}\zeta_i^2 = m \\ \mathsf{D}\eta &= 2m \end{aligned}$$

Утверждение. Пусть $\eta_m \sim \chi^2(m)$. Тогда, по ЦПТ,

$$\frac{\eta_m - \mathsf{E}\eta_m}{\sqrt{\mathsf{D}\eta_m}} = \frac{\eta_m - m}{\sqrt{2m}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Пример. $m = 50,\ \eta_m = 80$. Тогда

$$\frac{80 - 50}{10} = 3$$

и

$$\mathrm{cdf}_{\chi^2(50)}(80) = 0.9955 \approx \Phi(3) = 0.9986.$$

Предложение. $\chi^2(m)/m \xrightarrow[m\to\infty]{} 1$.

Доказательство. По ЗБЧ. □

[1] Вычисление $\mathsf{D}\eta$: https://www.statlect.com/probability-distributions/chi-square-distribution

3. Распределение Стьюдента $\mathrm{t}(m)$

Определение (Распределение $\mathrm{t}(m)$)**.** ξ имеет распределение Стьюдента с m степенями свободы, если

$$\xi \sim \mathrm{t}(m) \iff \xi = \frac{\zeta}{\sqrt{\eta/m}}, \quad \zeta \sim \mathrm{N}(0,1),\ \eta \sim \chi^2(m).$$

Свойства $\mathrm{t}(m)$

- При $m = 1$ это распределение Коши.
- При $m > 1$, $\mathsf{E}\xi = 0$ по симметричности.
- При $m > 2$, $\mathsf{D}\xi = m/(m-2)$.
- При $m > 3$, $\mathsf{A}\xi = 0$ по симметричности.
- При $m > 4$, $\mathsf{K}\xi = 6/(m-4)$.

Предложение. *Распределение Стьюдента сходится к стандартному нормальному:*

$$\mathrm{t} \Rightarrow \mathrm{N}(0,1).$$

Соображения по поводу. $\mathsf{D}\xi \to 1,\ \mathsf{K}\xi \to 0.$ □

4. Распределение Фишера

Определение. Распределение Фишера имеет вид

$$\mathrm{F}(m,k) = \frac{\chi^2(m)/m}{\chi^2(k)/k}.$$

Замечание. $\mathrm{F}(1,k) \sim \mathrm{t}^2(k)$; $\mathrm{F}(m,\infty) = \chi^2(m)/m$ потому что $\chi^2(k)/k \xrightarrow[k\to\infty]{} 1$.

5. Квадратичные формы от нормально распределенных случайных величин

(Это на след. семестр, сейчас можно не вникать.)
Пусть $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^\mathsf{T} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_p)$, $\mathbf{B}$ — симметричная, неотрицательно определенная матрица. Найдем распределение $\boldsymbol{\xi}^\mathsf{T}\mathbf{B}\boldsymbol{\xi}$.

Утверждение. Пусть $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_p)$, $\mathbf{B}, \mathbf{C}$ — симметричные матрицы размерности $p \times p$. Тогда $\boldsymbol{\xi}^\mathsf{T}\mathbf{B}\boldsymbol{\xi} \perp\!\!\!\perp \boldsymbol{\xi}^\mathsf{T}\mathbf{C}\boldsymbol{\xi} \iff \mathbf{BC} = \mathbf{0}$.

Пример (Независимость $\bar{x}^2$ и s^2)**.** Запишем

$$\bar{x}^2 = \frac{1}{n^2}\left(\sum_{i=1}^n x_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n x_i x_j = \frac{1}{n}\mathbf{x}\underbrace{\begin{pmatrix} 1/n & \dots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \dots & 1/n \end{pmatrix}}_{\mathbf{B}}\mathbf{x}^\mathsf{T}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n}\mathbf{x}\mathbf{B}\mathbf{x}^\mathsf{T} = \frac{1}{n}\left(\mathbf{x}\mathbf{I}_n\mathbf{x}^\mathsf{T} - \mathbf{x}\mathbf{B}\mathbf{x}^\mathsf{T}\right) = \frac{1}{n}\mathbf{x}\underbrace{\begin{pmatrix} 1-1/n & \dots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \dots & 1-1/n \end{pmatrix}}_{\mathbf{C}=\mathbf{I}_n-\mathbf{B}}\mathbf{x}^\mathsf{T}.$$

Таким образом, $n\bar{x}^2 = \mathbf{x}\mathbf{B}\mathbf{x}^\mathsf{T}$ и $ns^2 = \mathbf{x}\mathbf{C}\mathbf{x}^\mathsf{T}$. Но

$$\mathbf{BC} = \mathbf{B}(\mathbf{I}_n - \mathbf{B}) = \mathbf{B} - \mathbf{B}^2 = \mathbf{0},$$

так как

$$\mathbf{B}^2 = \begin{pmatrix} 1/n & \dots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \dots & 1/n \end{pmatrix}^2 = \begin{pmatrix} n\cdot 1/n & \dots & n\cdot 1/n \\ \vdots & \ddots & \vdots \\ n\cdot 1/n & \dots & n\cdot 1/n \end{pmatrix} = \mathbf{B}.$$

Значит, $\bar{x}^2 \perp\!\!\!\perp s^2$.

Видно, что $\sigma^{-2}\boldsymbol{\xi}^\mathsf{T}\mathbf{I}_p\boldsymbol{\xi} \sim \chi^2(p)$. На самом деле справедливо

Утверждение. Пусть $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_p)$, $\mathbf{B}$ — симметричная, неотрицательно неопределенная матрица размерности $p \times p$ и $\operatorname{rk}\mathbf{B} = r$. Тогда

$$\sigma^{-2}\boldsymbol{\xi}^\mathsf{T}\mathbf{B}\boldsymbol{\xi} \sim \chi^2(r) \iff \mathbf{B}^2 = \mathbf{B}.$$

Пример. Покажем, что

$$n\sigma^{-2}s^2 \sim \chi^2(p-1).$$

Воспользуемся представлением из предыдущего примера: $ps^2 = \mathbf{x}^\mathsf{T}\mathbf{C}\mathbf{x}$. Но $\operatorname{rk}\mathbf{C} = \operatorname{rk}(\mathbf{I}_p - \mathbf{B}) = p - 1$; $\mathbf{B}^2 = \mathbf{B}$, значит $p\sigma^{-2}s^2 \sim \chi^2(p-1)$.

Утверждение (Cochran). Пусть $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, $\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi} = \sum_i Q_i$, где Q_i — квадратичная форма, заданная $\mathbf{B}_i$, $\operatorname{rk}\mathbf{B}_i = r_i$. Тогда следующие утверждения эквивалентны:

1. $\sum r_i = p$
2. $Q_i \sim \chi^2(r_i)$
3. $Q_i \perp\!\!\!\perp Q_j, \quad \forall i \neq j$, т.е. $\mathbf{B}_i\mathbf{B}_j = \mathbf{0}$.

6. Распределение важных статистик

Пусть $\xi \sim \mathrm{N}(a, \sigma^2)$.

Предложение. $t = \sqrt{n}\frac{(\bar{x}-\mathsf{E}\xi)}{\sigma}$ *имеет стандартное нормальное распределение.*

Доказательство.

$$t = \frac{\bar{x}-a}{\sqrt{\mathsf{D}\bar{x}}} = \sqrt{n}\,\frac{\bar{x}-a}{\sigma} \sim \mathrm{N}(0,1).$$

□

Определим $s_a^2 = \sum_{i=1}^n (x_i - a)^2/n$.

Предложение. $ns_a^2/\sigma^2 \sim \chi^2(n)$.

Доказательство.

$$\chi^2 = \frac{ns_a^2}{\sigma^2} = \frac{n \cdot 1/n \cdot \sum_{i=1}^n (x_i - a)^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - a}{\sigma}\right)^2 \sim \chi^2(n).$$

□

Предложение. $ns^2/\sigma^2 = (n-1)\tilde{s}^2/\sigma^2 \sim \chi^2(n-1)$.

Доказательство. См. раздел 5). □

Альтернативное доказательство. По определению запишем

$$\underbrace{\mathsf{D}\hat{\xi}_n}_{s^2} = \mathsf{D}(\hat{\xi}_n - a) = \underbrace{\mathsf{E}(\hat{\xi}_n - a)^2}_{s_a^2} - \underbrace{(\mathsf{E}(\hat{\xi}_n - a))^2}_{(\bar{x}-a)^2}.$$

Домножив обе части на n/σ^2, получим

$$\frac{ns^2}{\sigma^2} = \frac{ns_a^2}{\sigma^2} - \frac{n(\bar{x}-a)^2}{\sigma^2} = \underbrace{\frac{ns_a^2}{\sigma^2}}_{\sim\chi^2(n)} - \underbrace{\left(\frac{\sqrt{n}\,(\bar{x}-a)}{\sigma}\right)^2}_{\sim\chi^2(1)} \Longrightarrow \frac{ns^2}{\sigma^2} \sim \chi^2(n-1).$$

□

Замечание. Для строгого доказательства, нужно использовать независимость $\bar{x}^2$ и s^2 (см. раздел 5).

Предложение. *Следующая статистика имеет распределение Стьюдента:*

$$t = \sqrt{n-1}\,\frac{\bar{x}-a}{s} = \frac{\sqrt{n-1}(\bar{x}-a)}{\sqrt{n-1}/\sqrt{n}\cdot\tilde{s}} = \sqrt{n}\,\frac{\bar{x}-a}{\tilde{s}} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1).$$

Предложение. $t = \sqrt{n-1}\,\frac{\bar{x}-a}{s} = \sqrt{n}\,\frac{\bar{x}-a}{\tilde{s}} \sim \mathrm{t}(n-1)$.

Доказательство.

$$t = \frac{\sqrt{n-1}\,(\bar{x}-a)}{s} = \frac{\sqrt{n-1}\left(\dfrac{\bar{x}-a}{\sigma}\right)}{s/\sigma} = \frac{\left(\dfrac{\bar{x}-a}{\sigma}\right)}{\sqrt{\dfrac{s^2/\sigma^2}{n-1}}} = \frac{\dfrac{\sqrt{n}(\bar{x}-a)}{\sigma}}{\sqrt{\dfrac{ns^2/\sigma^2}{n-1}}} = \frac{\zeta}{\sqrt{\eta/(n-1)}} \sim \mathrm{t}(n-1),$$

поскольку

$$\zeta = \frac{\sqrt{n}(\bar{x}-a)}{\sigma} \sim \mathrm{N}(0,1), \quad \eta = \frac{ns^2}{\sigma^2} \sim \chi^2(n-1).$$

и они независимы (также пока без доказательства — используется техника квадратичных форм или можно доказать через разложение дисперсии). □

Часть III.

Проверка гипотез и доверительные интервалы

Этот раздел иногда называется «Confirmatory Data Analysis» в противовес «Exploratory Data Analysis», не включающему в себя понятие *гипотезы*.

1. Построение критерия

Тут должно быть про определение критерия и построение его через статистику критерия.

2. Проверка гипотезы о значении параметра (характеристики)

2.1. Проверка гипотезы о значении мат. ожидания (t-критерий)

$H_0 : \mathsf{E}\xi = a = a_0$. Соответствие оценки математического ожидания гипотезе удобно выражать разницей $\bar{x} - a_0$ с «идеальным» значением 0. Отнормировав эту разницу, получим статистику, распределение которой известно.

2.1.1. $\mathsf{D}\xi = \sigma^2 < \infty$

Предложение. *Пусть $\mathsf{D}\xi = \sigma^2 < \infty$; тогда используется следующая статистика (z-score)*

$$t = z = \sqrt{n}\frac{(\bar{x} - a_0)}{\sigma} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1)$$

Предложение. *При условии $\xi \sim \mathrm{N}(a, \sigma^2)$,*

$$t = z \sim \mathrm{N}(0,1).$$

Доказательство.

$$z = \frac{\bar{x} - a_0}{\sqrt{\mathsf{D}\bar{x}}} = \sqrt{n}\,\frac{\bar{x} - a_0}{\sigma} \sim \mathrm{N}(0,1).$$

□

2.1.2. $\mathsf{D}\xi$ неизвестна

Предложение. *Пусть $\mathsf{D}\xi$ неизвестна; тогда используется следующая статистика*

$$t = \sqrt{n-1}\,\frac{\bar{x} - a_0}{s} = \sqrt{n}\,\frac{\bar{x} - a_0}{\tilde{s}} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1).$$

Предложение. *При условии нормальности данных,*

$$t \sim \mathrm{t}(n-1).$$

2.1.3. z-критерий для пропорции в модели Бернулли

Пусть $\xi \sim \mathrm{Ber}(p)$. Поскольку $\mathsf{E}\xi = p$, можно воспользоваться только что введенной статистикой; учитывая $\mathsf{D}\xi = p(1-p)$,

$$T = \sqrt{n}\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Разница с общим случаем состоит в том, что в параметрической модели с одним параметром не нужно оценивать дисперсию. Так как все выражается через этот параметр, то имеет формулу для дисперсии через значение параметра, предполагаемое в нулевой гипотезе.

2.2. Проверка гипотезы о значении дисперсии в нормальной модели (критерий χ^2)

Пусть $\xi \sim \mathrm{N}(a, \sigma^2)$. $H_0 : \mathsf{D}\xi = \sigma^2 = \sigma_0^2$. Соответствие оценки дисперсии гипотезе удобно выражать отношением s^2/σ_0^2 (или s_a/σ_0^2 если a известно) с «идеальным» значением 1. Домножив на n, получим статистику, распределение которой известно.

2.2.1. $\mathsf{E}\xi = a < \infty$

Предложение. *Пусть* $\mathsf{E}\xi = a < \infty$*; При условии нормальности данных используется следующая статистика:*

$$\chi^2 = n\frac{s_a^2}{\sigma_0^2} \sim \chi^2(n).$$

2.2.2. $\mathsf{E}\xi$ неизвестно

Предложение. *Пусть* $\mathsf{E}\xi$ *неизвестно. При условии нормальности данных используется следующая статистика:*

$$\chi^2 = n\frac{s^2}{\sigma_0^2} = (n-1)\frac{\tilde{s}^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Упражнение. $s^2 = 1.44, \bar{x} = 55, n = 101$. Проверить гипотезу $\sigma_0^2 = 1.5$ в нормальной модели.

Решение. Воспользуемся статистикой

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = 101 \cdot 0.96 = 96.96.$$

«Идеальные» значения близки к $\mathsf{E}\xi_{\chi^2(100)} = 100$, так что определим критическую область на концах плотности:

$$p\text{-value}/2 = \mathrm{cdf}_{\chi^2(100)}(96.96) = \texttt{pchisq}(\texttt{96.96}, \texttt{100}) \approx 0.43 \implies p\text{-value} \approx 0.86.$$

Замечание. Можно посчитать и по таблицам для нормального распределения. Раз

$$\frac{\eta_m - \mathsf{E}\eta_m}{\sqrt{\mathsf{D}\eta_m}} \xrightarrow[m\to\infty]{\mathrm{d}} \mathrm{N}(0,1),$$

то

$$\frac{96.96 - 100}{\sqrt{200}} \approx -0.215 \implies p\text{-value}/2 = \Phi(-0.215) \approx 0.415.$$

⌟

3. Доверительные интервалы

3.1. Мотивация и определение

Точечные оценки не дают информации о том, насколько (количественно) настоящее значение далеко от оценки.

Определение. $[b_1, b_2]$ — *доверительный интервал* для параметра θ с уровнем доверия $\gamma \in [0,1]$, если $\forall \theta$

$$\mathsf{P}(\theta \in [b_1, b_2]) = \gamma, \quad \text{где } b_1 = b_1(\mathbf{x}), b_2 = b_2(\mathbf{x}) \text{ — статистики.}$$

Замечание. Если выборка из дискретного распределения, то b_1, b_2 — тоже дискретны. Поэтому наперед заданную точность получить может не получиться; в таких случаях знак «=» заменяют «$\geq$». Аналогично с заменой на «$\xrightarrow[n\to\infty]{}$» для асимптотических доверительных интервалов, когда точные получить невозможно или трудоемко.

3.2. Доверительный интервал для проверки гипотезы о значении параметра

Зафиксируем $H_0 : \theta = \theta_0$ и $\gamma = 1 - \alpha$, где α играет роль уровня значимости. По определению доверительного интервала, $\mathsf{P}(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \gamma$. Тогда

$$\mathsf{P}(\theta \in [b_1(\mathbf{x}), b_2(\mathbf{x})]) = \gamma = 1 - \alpha \implies \alpha = 1 - \mathsf{P}(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \mathsf{P}(\theta \notin [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})])$$

и $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus [b_1(\mathbf{x}), b_2(\mathbf{x})]$. Соответственно,

$$\begin{cases} \text{отвергаем } H_0, \text{ если} & \theta_0 \notin [b_1(\mathbf{x}), b_2(\mathbf{x})] \\ \text{не отвергаем } H_0, \text{ если} & \theta_0 \in [b_1(\mathbf{x}), b_2(\mathbf{x})]. \end{cases}$$

Вероятность ошибки первого рода равно α, что соответствует определению критерия. Заметим, что здесь мы пользуемся общим определением критерия, а не частным случаем, когда критерий строится через статистику критерия.

3.3. Доверительные интервалы для математического ожидания и дисперсии в нормальной модели

Предположение. *Пусть* $\xi \sim \mathrm{N}(a, \sigma^2)$.

3.3.1. Доверительный интервал для a

- Пусть σ^2 известно. Свяжем a с выборкой через статистику критерия $t = \sqrt{n}\frac{(\bar{x}-a)}{\sigma} \sim \mathrm{N}(0,1)$:

$$\gamma = \mathsf{P}(c_1 < t < c_2) = \mathsf{P}\left(c_1 < \sqrt{n}\frac{(\bar{x}-a)}{\sigma} < c_2\right) = \mathsf{P}\left(a \in \left(\bar{x} - \frac{\sigma c_2}{\sqrt{n}}, \bar{x} - \frac{\sigma c_1}{\sqrt{n}}\right)\right).$$

  Решений уравнения $\mathsf{P}\left(c_1 < \sqrt{n}(\bar{x}-a)/\sigma < c_2\right) = \Phi(c_2) - \Phi(c_1) = \gamma$ бесконечно много. Чем $[c_1, c_2]$ короче, тем лучше. Поскольку Φ симметрична и унимодальна,

$$\begin{matrix} c_1 = -c_\gamma \\ c_2 = c_\gamma, \end{matrix} \quad \text{где } c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\gamma + \frac{1-\gamma}{2}\right) = x_{\frac{1+\gamma}{2}}.$$

Наконец,

$$\mathsf{P}\left(a \in \left(\bar{x} \pm \frac{\sigma}{\sqrt{n}} c_\gamma\right)\right) = \gamma.$$

- Пусть σ^2 неизвестно. По аналогии,

$$\gamma = \mathsf{P}\left(c_1 < \frac{\sqrt{n-1}(\bar{x}-a)}{s} < c_2\right) = \mathsf{P}\left(a \in \left(\bar{x} \pm \frac{c_\gamma s}{\sqrt{n-1}}\right)\right), \quad c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{t}(n-1)}\left(\frac{1+\gamma}{2}\right)$$

и

$$\mathsf{P}\left(a \in \left(\bar{x} \pm \frac{\tilde{s}}{\sqrt{n}} c_\gamma\right)\right) = \gamma.$$

Упражнение. Пусть $s^2 = 1.21, \bar{x} = 1.9, n = 36$. Построить 95% доверительный интервал для $\mathsf{E}\xi$.

Решение.

$$c_\gamma = \mathtt{qt}(0.975, 35) \approx 2.03 \implies \left(1.9 \pm \frac{2.03 \cdot \sqrt{1.21}}{\sqrt{35}}\right) = (1.52; 2.28)\,.$$

⌟

3.3.2. Доверительный интервал для σ^2

- Пусть a известно. Поскольку плотность χ^2 становится все более симметричной с ростом n, примем

$$c_1 = \mathrm{cdf}^{-1}_{\chi^2(n)}\left(\frac{1-\gamma}{2}\right),\ c_2 = \mathrm{cdf}^{-1}_{\chi^2(n)}\left(\frac{1+\gamma}{2}\right).$$

Тогда

$$\mathsf{P}\left(c_1 < \frac{n s_a^2}{\sigma^2} < c_2\right) = \gamma \iff \mathsf{P}\left(\sigma^2 \in \left(\frac{n s_a^2}{c_2}, \frac{n s_a^2}{c_1}\right)\right) = \gamma.$$

- Пусть a неизвестно. Тогда аналогично

$$\mathsf{P}\left(\sigma^2 \in \left(\frac{n s^2}{c_2}, \frac{n s^2}{c_1}\right)\right) = \gamma,$$

где

$$c_1 = \mathrm{cdf}^{-1}_{\chi^2(n-1)}\left(\frac{1-\gamma}{2}\right),\ c_2 = \mathrm{cdf}^{-1}_{\chi^2(n-1)}\left(\frac{1+\gamma}{2}\right).$$

3.4. Асимптотический доверительный интервал для математического ожидания в модели с конечной дисперсией

Если модель неизвестна, но известно, что $\mathsf{D}\xi < \infty$, можно построить доверительный интервал для $\mathsf{E}\xi = a$. Пусть $\{x_i\}$ i.i.d., тогда

$$t = \frac{\sqrt{n}\,(\bar{x}-a)}{\sigma} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1).$$

Если положить $\sigma := s$, то сходимость не испортится, потому что s^2 — состоятельная оценка σ^2. Тогда

$$\mathsf{P}\left(\mathsf{E}\xi \in \left(\bar{x} \pm \frac{s c_\gamma}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} \gamma, \quad c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{t}(n-1)}\left(\frac{1+\gamma}{2}\right).$$

Альтернативно замену σ на s можно обосновать по теореме Слуцкого.

Утверждение (Слуцкий). Если $\xi_n \xrightarrow{\mathrm{d}} \xi,\ \eta_n \xrightarrow{\mathsf{P}} c$, то $\xi_n + \eta_n \xrightarrow{\mathrm{d}} \xi + c$ и $\xi_n \eta_n \xrightarrow{\mathrm{d}} c\xi$.

Используя тот факт, что $s \xrightarrow{\mathsf{P}} \sigma$, запишем

$$\mathsf{P}\left(c_1 < \frac{\sqrt{n}\,(\bar{x}-a)}{\sigma}\frac{\sigma}{s} < c_2\right) \xrightarrow[n\to\infty]{} \Phi(c_2) - \Phi(c_1).$$

Пример. Доверительный интервал для параметра $\mathrm{Exp}(\lambda)$. FIXME

3.5. Асимптотический доверительный интервал для параметра на основе MLE

Если умеем находить $\hat\theta_{\text{MLE}}$, т,о по асимптотической нормальности,

$$\frac{\hat\theta_{\text{MLE}} - \mathsf{E}\hat\theta_{\text{MLE}}}{\sqrt{\mathsf{D}\hat\theta_{\text{MLE}}}} \xrightarrow{\text{d}} \text{N}(0,1),$$

по асимптотической несмещенности,

$$\frac{\hat\theta_{\text{MLE}} - \theta}{\sqrt{\mathsf{D}\hat\theta_{\text{MLE}}}} \xrightarrow{\text{d}} \text{N}(0,1),$$

и, учитывая асимптотическую эффективность ($\mathsf{D}\hat\theta_{\text{MLE}} I_n(\theta) \xrightarrow[n\to\infty]{} 1$), запишем статистику

$$T = \left(\hat\theta_{\text{MLE}} - \theta\right)\sqrt{I_n(\theta)} \xrightarrow{\text{d}} \text{N}(0,1).$$

Чтобы по аналогии с предыдущим выразить θ в $\mathsf{P}(c_1 < T < c_2) = \mathsf{P}(|T| < c_\gamma) = \gamma$, необходимо выразить θ из $I_n(\theta)$. Для Pois и Ber это эквивалентно решению квадратного уравнения.

В общем случае, можно вместо θ в $I_n(\theta)$ подставить $\hat\theta_{\text{MLE}}$ (при $n \to \infty$ это не должно сильно испортить дело), откуда

$$\mathsf{P}(|T| < c_\gamma) = \gamma \iff \mathsf{P}\left(-c_\gamma < \left(\hat\theta_{\text{MLE}} - \theta\right)\sqrt{I_n(\theta)} < c_\gamma\right) = \gamma \iff \mathsf{P}\left(\theta \in \left(\hat\theta_{\text{MLE}} \pm \frac{c_\gamma}{\sqrt{I_n(\theta)}}\right)\right) = \gamma,$$

где

$$T \xrightarrow{\text{d}} \text{N}(0,1) \implies c_\gamma = \text{cdf}^{-1}_{\text{N}(0,1)}\left(\frac{1+\gamma}{2}\right).$$

Пример. $\xi \sim \text{Pois}(\lambda)$. По 5.3, $\hat\lambda_{\text{MLE}} = \bar x$, по 6.4.1 $I_n(\lambda) = n/\lambda = n/\bar x$ откуда

$$\mathsf{P}\left(\lambda \in \left(\bar x \pm \text{cdf}^{-1}_{\text{N}(0,1)}\left(\frac{1+\gamma}{2}\right)\frac{\sqrt{\bar x}}{\sqrt{n}}\right)\right) = \gamma.$$

Пример. $\xi \sim \text{Ber}(p)$. $p = \mathsf{E}\xi$. $\hat p = \bar x$, откуда

$$\mathsf{P}\left(p \in \left(\hat p \pm c_\gamma \frac{\sqrt{\hat p(1-\hat p)}}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} \gamma.$$

Замечание. Этот доверительный интервал не очень хорош, потому что не обязательно принадлежит $[0,1]$.

3.6. Использование SE для построения доверительных интервалов

Пусть оценка $\hat\theta_n$ имеет какое-то симметричное распределение хотя бы асимптотически. Как и для любой другой случайной величины (с симметричным распределением), доверительный интервал уровня γ (т.е. такой интервал, в котором лежит γ всех значений величины) задается как

$$\mathsf{E}\hat\theta_n \pm c_\gamma\sqrt{\mathsf{D}\hat\theta_n},$$

где $c_\gamma = \text{qnt}\,\gamma$. К примеру, для $\text{N}(0,1)$ и 95%-квантили это был бы интервал$(-1.96; 1.96)$, а так нужно передвинуть его на среднее и растянуть на корень из дисперсии.

Но стандартное отклонение $\sqrt{\mathsf{D}\hat\theta_n}$ распределения $\hat\theta_n$ можно оценить как SE. Значит доверительный интервал будет иметь вид

$$\mathsf{E}\hat\theta_n \pm c_\gamma\,\text{SE}\,.$$

4. Критерии согласия с видом распределения

4.1. Критерий χ^2 согласия с видом распределения

По выборке возможно проверить гипотезу о виде распределения случайной величины, реализацией которой является выборка.

Утверждение. Для проверки гипотезы согласия с видом произвольного *дискретного* распределения используется асимптотический *критерий* χ^2 *(«chi-squared test for goodness of fit»).*

4.1.1. Распределение с известными параметрами

Пусть

$$H_0 : \mathcal{P} = \mathcal{P}_0, \text{ где } \mathcal{P}_0 : \begin{pmatrix} x_1^* & \dots & x_k^* \\ p_1 & \dots & p_k \end{pmatrix}.$$

Сгруппируем $\mathbf{x}$; каждому x_i^* сопоставим *эмпирическую* абсолютную частоту ν_i; тогда np_i — *ожидаемая* абсолютная частота.

В качестве меры расхождения между эмпирическим и генеральным распределением рассматривается величина

$$\sum_{i=1}^{k} c_i \left(\frac{\nu_i}{n} - p_i \right)^2, \quad c_i = \frac{n}{p_i},$$

откуда записывается статистика критерия

$$T = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$$

с идеальным значением 0.

Утверждение. $T \xrightarrow{\text{d}} \chi^2(k-1)$.

Определение. Критерий *применим*, если $\alpha_{\text{I}} = \alpha$ или $\alpha_{\text{I}} \approx \alpha$ с достаточной степенью точности.

Замечание. Поскольку критерий асимптотический, с достаточной степенью точностью он применим в случае, если

1. $n \geq 50$;
2. $np_i \geq 5$.

Замечание. Если условие $np_i \geq 5$ не выполняется, следует объединить состояния, например, с краев или слева направо; если в хвосте оказалось < 5, то следует присоединить к последнему.

Замечание. Почему бы не подстраховаться и не объединить состояния так, чтобы было >10? Ответ: теряем в мощности.

Пример (С монеткой)**.** Пусть $n = 4040$, $\#\text{H} = 2048$, $\#\text{T} = 1092$. Проверим $H_0 : \mathcal{P} = \text{Ber}(0.5)$ с $\alpha = 0.1$. Условия критерия выполняются, поэтому посчитаем

$$T = \frac{(2048 - 2020)^2}{2020} + \frac{(1092 - 2020)^2}{2020} = \frac{28^2 + 28^2}{2020} \approx 0.78 \sim \chi^2(1),$$

откуда

$$p\text{-value} = 1 - \mathrm{cdf}_{\chi^2(1)}(0.78) \approx 0.38.$$

$0.38 > 0.1$, значит H_0 не отвергается.

Замечание. Прохождение критерия не достаточно. Так, альтернирующая (и явно не случайная) последовательность $\mathbf{x} = (0, 1, 0, 1, \ldots)$ имеет $T = 0$.

4.1.2. Распределение с неизвестными параметрами

В случае сложной гипотезы $\mathcal{P} \in \{\mathcal{P}(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^\mathsf{T}$, следует найти оценку $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ (или $\hat{\boldsymbol{\theta}} : \hat{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$) по методу максимального правдоподобия. При подстановке оценок вместо истинных параметров критерий становится консервативным. Чтобы этого избежать, необходимо сделать поправку на количество параметров — отнять r. Что приятно, одна и та же поправка работает для всех распределений; в этом случае,

$$T \xrightarrow{\mathrm{d}} \chi^2(k - r - 1).$$

5. Критерий Колмогорова-Смирнова согласия с видом распределения

5.1. Произвольное абсолютно непрерывное распределение

$H_0 : \xi \sim \mathcal{P} = \mathcal{P}_0$.

Утверждение. Для проверки гипотезы согласия с видом произвольного *абсолютно непрерывного* распределения с известными параметрами используется асимптотический критерий Колмогорова-Смирнова со следующей статистикой:

$$D_n = \sup_{x \in \mathbf{x}} \left| \widehat{\mathrm{cdf}}_n(x) - \mathrm{cdf}_0(x) \right|,$$

где cdf_0 — функция распределения $\mathcal{P}_0$ нулевой гипотезы.

Альтернатива только одна: $H_1 : \xi \not\sim \mathcal{P}_0$; $\mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\text{K-S}}(1-\alpha), \infty)$.

Замечание. Критерий является *точным*, не асимптотическим. Значит, им можно пользоваться и при маленьких объемах выборки (мощность, при этом, останется низкой все-равно).

Замечание. $\sup_x \sqrt{n} \left| \widehat{\mathrm{cdf}}_n(x) - \mathrm{cdf}_0(x) \right| \xrightarrow{\mathrm{d}} \mathcal{P}_{\text{K.S.}}$, где $\mathcal{P}_{\text{K.S.}}$ — распределение Колмогорова. Значит, при больших объемах выборки для такой статистики критерия можно пользоваться таблицами распределения Колмогорова.

6. Визуальное определение согласия с распределением

6.1. P-P plot

Определение. *P-P plot* есть график

$$\left\{\left(\mathrm{cdf}_0(x_i) + \frac{1}{2n}, \widehat{\mathrm{cdf}}_n(x_i)\right)\right\}_{i=1}^{n}.$$

Пример. В R:

```
pp.plot <- function(xs, cdf.0=pnorm, n.knots=1000) {
  knots <- seq(min(xs), max(xs), length.out=n.knots)
  plot(cdf.0(knots), ecdf(xs)(knots))
  abline(0, 1)
}
```

6.2. Q-Q plot

Определение. *Q-Q plot* есть график

$$\left\{\left(x_i, \mathrm{cdf}_0^{-1}\left(\widehat{\mathrm{cdf}}_n(x_i) + \frac{1}{2n}\right)\right)\right\}_{i=1}^{n}.$$

Определение. Частный случай Q-Q plot для $\mathrm{cdf}_0^{-1} = \mathrm{cdf}_{\mathrm{N}(0,1)}^{-1}$ называется *normal probability plot*.

Пример. В R:

```
qq.plot <- function(xs, qf.0=qnorm, n.ppoints=1000) {
  qs <- ppoints(n.ppoints)
  plot(qf.0(qs), unname(quantile(xs, probs=qs)))
  abline(mean(xs), sd(xs))
}
```

Замечание. Если $\hat{\mathcal{P}}_n \to \mathcal{P}_\xi$, то оба графика будут стремиться к $y = x$. Референсной прямой normal probability plot будет $y = \widehat{\mathsf{D}\xi} \cdot x + \widehat{\mathsf{E}\xi}$.

Замечание. Больше о различии Q-Q и P-P plots, см. http://v8doc.sas.com/sashtml/qc/chap8/sect9.htm

Замечание. Различные интерпретации параметров распределения по Q-Q plot можно посмотреть в интерактивном приложении: https://xiongge.shinyapps.io/QQplots/

Часть IV.

Корреляционный анализ

Определение. *Мера зависимости* — это функционал $r : (\xi, \eta) \mapsto x \in [-1, 1]$ со свойствами:

1. $|r| \leq 1$.
2. $\xi \perp\!\!\!\perp \eta \implies r(\xi, \eta) = 0$.
3. Если ξ и η «максимально зависимы», то $r(\xi, \eta) = 1$.

1. Вероятностная независимость

1.1. Визуальное определение независимости

- Поскольку при $p_\eta(y_0) \neq 0$

$$\xi \perp\!\!\!\perp \eta \iff p_{\xi|\eta}(x \mid y_0) = \frac{p_{\xi,\eta}(x, y_0)}{p_\eta(y_0)} = p_\xi(x),$$

  то срезы графика совместной плотности при фиксированном y_0 после нормировки $p_\eta(y_0)$ должны выглядеть одинаково для всех y_0.

- Для выборки независимость можно попытаться определить по *таблицам сопряженности*: сгруппируем $\{(x_i, y_i)\}_{i=1}^n$ и сопоставим каждой уникальной паре абсолютную частоту ν_{ij}:

$$\begin{array}{cccc} & y_1^* & \cdots & y_s^* \\ x_1^* & \nu_{11} & \cdots & \nu_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^* & \nu_{k1} & \cdots & \nu_{ks} \end{array}$$

  Тогда признаки с большей чем случайной вероятностью будут независимы при пропорциональных строчках / столбцах. Более формально, признаки независимы, если

$$\frac{\nu_{ij}}{\sum_k \nu_{kj}} = \frac{\nu_{ij}}{\nu_{\cdot j}} = \hat{p}_{i|j} \propto \hat{p}_{i|\ell},$$

  т.е. вероятности условного распределения не зависят от выбора строки.

  Пример. Таблица сопряженности похожей на независимую выборки:

$$\begin{array}{ccc} 1 & 3 & 2 \\ 2 & 5 & 3 \\ 9 & 20 & 11 \end{array}$$

1.2. Критерий независимости χ^2

По определению, для двумерных дискретных распределений, независимость есть

$$\xi \perp\!\!\!\perp \eta \iff \underbrace{\mathsf{P}(\xi = i, \eta = j)}_{p_{ij}} = \underbrace{\mathsf{P}(\xi = i)}_{p_i}\underbrace{\mathsf{P}(\eta = j)}_{p_j} = \underbrace{\sum_{k=1}^{K} \mathsf{P}(\xi = i, \eta = k)}_{p_{i\cdot}} \cdot \underbrace{\sum_{s=1}^{S} \mathsf{P}(\xi = s, \eta = j)}_{p_{\cdot j}}.$$

Проверим $H_0 : \xi \perp\!\!\!\perp \eta$.

Утверждение. ОМП оценкой будет $\hat{p}_{i\cdot} = \nu_{i\cdot}/n$ и $\hat{p}_{\cdot j} = \nu_{\cdot j}/n$.

Следовательно,

$$\xi \perp\!\!\!\perp \eta \iff \hat{p}_{ij} = \frac{\nu_{ij}}{n} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{\nu_{i\cdot}}{n} \cdot \frac{\nu_{\cdot j}}{n}.$$

Это равенство удается получить редко; важно определить, не является ли это нарушение случайным.

Запишем статистику

$$\chi^2 = \sum_{i=1}^{K}\sum_{j=1}^{S} \frac{(\nu_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_{i=1}^{K}\sum_{j=1}^{S} \frac{(\nu_{ij} - \nu_{i\cdot}\nu_{\cdot j}/n)^2}{\nu_{i\cdot}\nu_{\cdot j}/n} \xrightarrow{\mathrm{d}} \chi^2((k-1)(s-1))$$

Количество параметров таково, потому что если $\xi \parallel \eta$, то всего $ks - 1$ параметров (-1 потому что $\sum_{ij} p_{ij} = 1$); если $\xi \perp\!\!\!\perp \eta$, то $k + s - 2$ (-2 потому что $\sum_i p_{ij} = 1$ и $\sum_j p_{ij} = 1$). Значит $ks - 1 - k - s + 2 = (k-1)(s-1)$.

Пример. Дано S кубиков. Проверить гипотезу, что кубики одинаковы.

Решение. FIXME ⌟

Замечание. На маленьких выборках ($n < 40$, $np_{ij} < 5$) возникают проблемы со сходимостью, потому что можно объединять только столбцы / строки и каждый раз терять сразу $S-1$ ($K-1$) степень свободы. В этих случаях используют критерием с перестановкой[1] или, в случае таблиц сопряженности 2×2, точным критерием Фишера.

Замечание. Критерий верен для количественных, порядковых и качественных признаков, потому что нигде не участвуют значения из выборки.

Замечание. Критерий асимптотический, поэтому $\alpha_{\mathrm{I}} \to \alpha$.

Замечание. Критерий не удовлетворяет 1-му пункту определения меры зависимости ($\chi^2 \notin [-1, 1]$). Это обычно исправляют так: рассматривают *среднеквадратичную сопряженность*

$$r^2 := \frac{\chi^2}{n}$$

или коэффициент сопряженности Пирсона

$$p^2 := \frac{\chi^2}{\chi^2 + n}$$

(тогда 1 никогда не достигается). Могли бы работать с $1 - p$-value, но так почему-то никогда не делают.

[1] https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests

2. Линейная / полиномиальная зависимость

Пусть теперь ξ, η — количественные признаки.

Определение. Определим

$$\phi(x) := \mathsf{E}\left\{\eta \mid \xi = x\right\}.$$

Тогда назовем зависимость *линейной*, если $\phi(x)$ — линейная функция, *квадратичной* — если квадратичная и т.д.

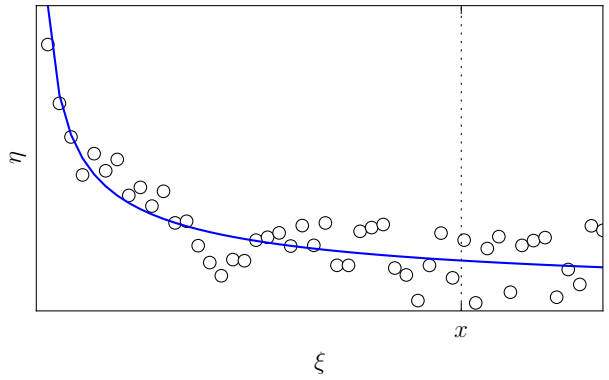

Рис. 2.1.: Нелинейная зависимость

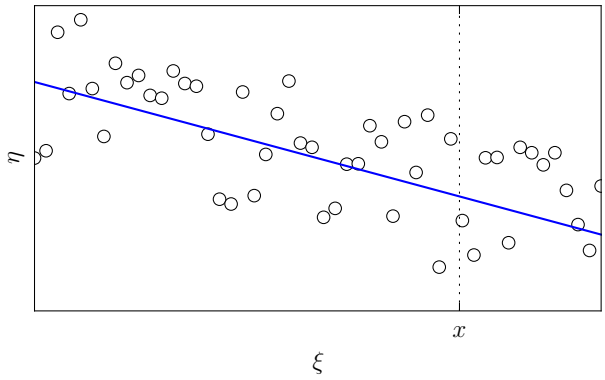

Рис. 2.2.: Линейная зависимость

Определение. Мера *линейной* зависимости между случайным величинами ξ и η есть *коэффициент корреляции Пирсона*

$$\rho = \frac{\operatorname{cov}(\xi, \eta)}{\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}\eta}}.$$

Замечание. Про ρ можно думать как про cos между векторами в соответствующем пространстве.

Замечание (Важное).

$$\begin{aligned} \xi \perp\!\!\!\perp \eta \quad &\Longrightarrow \quad \rho = 0 \\ \xi, \eta \sim \mathrm{N}(\mu, \sigma^2),\ \xi \perp\!\!\!\perp \eta \quad &\Longleftrightarrow \quad \rho = 0. \end{aligned}$$

Предложение. *Для линейно зависимых данных, конечно,* $\rho = 1$.

Доказательство. Пусть $\eta = a + b\xi$; тогда

$$\begin{aligned} \rho(\xi, \eta) &= \frac{\operatorname{cov}(\xi, a + b\xi)}{\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}(a + b\xi)}} = \frac{\mathsf{E}\xi(a + b\xi) - \mathsf{E}\xi\mathsf{E}(a + b\xi)}{\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}b\xi}} = \frac{\mathsf{E}\xi a + b\mathsf{E}\xi^2 - \mathsf{E}\xi\mathsf{E}a - \mathsf{E}\xi b\mathsf{E}\xi}{b\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}\xi}} = \\ &= \frac{a\mathsf{E}\xi + b\mathsf{E}\xi^2 - a\mathsf{E}\xi - b(\mathsf{E}\xi)^2}{b\mathsf{D}\xi} = \frac{b(\mathsf{E}\xi^2 - (\mathsf{E}\xi)^2)}{b\mathsf{D}\xi} = 1. \end{aligned}$$

□

2.1. О соотношении ρ и коэффициента линейной регрессии

По (2), если линейная регрессия уравнением $y = kx + b$, то

$$k = \rho\frac{\sigma_\eta}{\sigma_\xi}.$$

В общем случае, по виду прямой линейной регрессии ничего нельзя сказать о зависимости между случайными величинами. Так, если $\eta = a + b\xi$ есть линейная функция от ξ, то, по предыдущему, $\rho = 1$ и

$$k = 1 \cdot \frac{\sqrt{\mathsf{D}(a + b\xi)}}{\sqrt{\mathsf{D}\xi}} = b$$

и прямая может иметь произвольный, в зависимости от b, наклон.

Замечание. В то же время, поскольку для

$$\begin{pmatrix}\xi\\ \eta\end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma), \quad \Sigma = \begin{pmatrix}\sigma_\xi^2 & \operatorname{cov}(\xi,\eta)\\ \operatorname{cov}(\xi,\eta) & \sigma_\eta^2\end{pmatrix}$$

справедливо, что

$$k = \rho\frac{\sigma_\eta}{\sigma_\xi} = \frac{\operatorname{cov}(\xi,\eta)}{\sigma_\xi\sigma_\eta} \cdot \frac{\sigma_\eta}{\sigma_\xi} = \frac{1}{\sigma_\xi^2}\operatorname{cov}(\xi,\eta),$$

то $k = 0 \iff \operatorname{cov}(\xi,\eta) = 0$, а для стандартно нормальных данных $k = \rho = \operatorname{cov}(\xi,\eta)$.

2.2. Значимость коэффициента корреляции

Определение. Коэффициент корреляции *значим*, если отвергается $H_0 : \rho = 0$.

Чаще, чем $H_0 : \rho = \rho_0$, проверяют $H_0 : \rho > \rho_0$. Если $\rho_0 = 0$, то $(\xi,\eta)^\mathsf{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ и, по ЦПТ,

$$T = \frac{\sqrt{n-2}\hat\rho_n}{\sqrt{1-\hat\rho_n^2}} \sim \mathrm{t}(n-2).$$

Идеальное значение — 0, два хвоста.

Если $\rho_0 \neq 0$, то ЦПТ не работает, тогда распределение $\hat\rho$ неизвестно. Тогда применяется z-преобразование Фишера

$$z = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \quad z_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}.$$

Тогда ЦПТ работает и, если $(\xi,\eta)^\mathsf{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$T = \sqrt{n-3}(z - z_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

3. Метод наименьших квадратов (Ordinary Least Squares)

Пусть $\eta, \xi \in L^2(\mathcal{F}, \mathsf{P})$ пространству $\mathcal{F}$-измеримых по мере P функций с конечным вторым моментом и скалярным произведением $(\eta, \xi) = \mathsf{E}\eta\xi$, причем $\hat{\eta} \in K = \{\phi(\xi)\} = \{\hat{\eta} : \sigma(\phi(\xi))\text{-измерима}\}$. По свойству УМО(6.2), вектор

$$\hat{\eta}^* = \mathsf{E}\{\eta \mid \phi(\xi)\}$$

будет ортогональной проекцией η на K, т.е. $(\eta - \hat{\eta}^*, \hat{\eta}) = 0 \ \forall \hat{\eta} \in K$. Значит, он минимизирует квадрат нормы расстояния от η до K:

$$\hat{\eta}^* = \operatorname*{argmin}_{\hat{\eta} \in K} \|\eta - \hat{\eta}\|^2 = \operatorname*{argmin}_{\hat{\eta} \in K} \mathsf{E}(\eta - \hat{\eta})^2 = \mathsf{E}\{\eta \mid \phi(\xi)\}.$$

$\hat{\eta}^*$ называется *наилучшим среднеквадратичным приближением в классе* K.

4. Корреляционное отношение

Если $K = \mathcal{L} = \{a\xi + b\}$ — линейное пространство, то теорема Пифагора принимает вид

$$\mathsf{D}\eta = \mathsf{E}\,(\eta - \mathsf{E}\eta)^2 = \underbrace{\mathsf{E}\,(\hat{\eta}^* - \mathsf{E}\eta)^2}_{\text{объяснённая доля аппроксимации}} + \underbrace{\mathsf{E}\,(\eta - \hat{\eta}^*)^2}_{\text{ошибка аппроксимации}}.$$

Откуда можно записать меру аппроксимации

$$\frac{\mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2}{\mathsf{D}\eta} = 1 - \frac{\mathsf{E}\,(\eta - \hat{\eta}^*)^2}{\mathsf{D}\eta} = 1 - \frac{\min_{\hat{\eta}\in\mathcal{L}} \mathsf{E}\,(\eta - \hat{\eta})^2}{\mathsf{D}\eta}.$$

Определение. Полученная величина называется коэффициентом корреляции ρ^2:

$$\rho^2 := 1 - \frac{\min_{\hat{\eta}\in\mathcal{L}} \mathsf{E}\,(\eta - \hat{\eta})^2}{\mathsf{D}\eta}.$$

ρ — коэффициент корреляции Пирсона.

Определение. *Множественный коэффициент корреляции* есть полученная величина для МНК с $K = \mathcal{M} = \left\{\sum_{i=1}^k b_i\xi_i + b_0\right\}$.

$$R^2(\eta, \xi_1, \ldots, \xi_k) := 1 - \frac{\min_{\hat{\eta}\in\mathcal{M}} \mathsf{E}\,(\eta - \hat{\eta})^2}{\mathsf{D}\eta}.$$

Замечание. $R^2 \geq \rho^2$; если же $R^2 = \rho^2$, то $\xi_1, \ldots, \xi_k$ все зависимы.

Определение. В общем случае, если $K = \{\phi(\xi)$ измеримые$\}$, то полученная величина называется *корреляционным отношением*:

$$r^2_{\eta|\xi} := 1 - \frac{\min_{\hat{\eta}\in K} \mathsf{E}\,(\eta - \hat{\eta})^2}{\mathsf{D}\eta} = \frac{\mathsf{DE}(\eta \mid \xi)}{\mathsf{D}\eta}.$$

4.1. Свойства корреляционного отношения

1. $r^2_{\eta|\xi} \in [0, 1]$.
2. $\eta \perp\!\!\!\perp \xi \implies r^2_{\eta|\xi} = 0$.
3. $\eta = \phi(\xi) \iff r^2_{\eta|\xi} = 1$.
4. Вообще говоря, $r^2_{\eta|\xi} \neq r^2_{\xi|\eta}$. К примеру, для любой не монотонной функции (так, чтобы не существовала обратная).
5. $r^2_{\eta|\xi} \geq \rho^2(\eta, \xi)$ (потому что минимум по всем функциям меньше, чем лишь по линейным, значит $1 - \min$ больше).
6. $(\xi, \eta)^\mathsf{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies r^2_{\eta|\xi} = \rho^2(\eta, \xi)$.

4.2. Выборочное корреляционное отношение

По разложению дисперсии,

$$\mathsf{D}\eta = \mathsf{E}(\eta - \mathsf{E}\eta)^2 = \underbrace{\mathsf{E}(\mathsf{E}(\eta \mid \xi) - \mathsf{E}\eta)^2}_{\mathsf{DE}(\eta|\xi)} + \mathsf{E}(\eta - \mathsf{E}(\eta \mid \xi))^2.$$

Перейдем на выборочный язык. Пусть дана выборка

$$\binom{x_1}{y_1}, \ldots, \binom{x_n}{y_n}.$$

Сгруппируем её:

$$\begin{array}{c|ccc} x_1^* & y_{11} & \ldots & y_{1n_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^* & y_{k1} & \ldots & y_{kn_k} \end{array}$$

Пусть ξ — дискретная случайная величина со значениями $(x_1^*, \ldots, x_k^*)$. Тогда, учитывая

$$\bar{y}_i = \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_1} y_{ij} = \hat{\mathsf{E}}(\eta \mid \xi = x_i^*),$$

на выборочном языке получаем (домножив на n):

$$\begin{aligned} \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}_{\text{total sum of squares}} &= \underbrace{\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2}_{\text{межгрупповой разброс}} + \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}_{\text{внутригрупповой разброс}} \\ n s_y^2 &= n s_{y|x}^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \end{aligned}$$

Отсюда, так как, $r_{\eta|\xi}^2 = \mathsf{DE}(\eta \mid \xi)/\mathsf{D}\eta$,

$$\hat{r}_{\eta|\xi}^2 = \hat{r}_{y|x}^2 = \frac{s_{y|x}^2}{s_y^2}.$$

5. Частная корреляция

Определение. Частная корреляция случайных величин η_1, η_2 относительно $\{\xi_1, \dots, \xi_k\}$ есть

$$\rho\left(\eta_1, \eta_2 \mid \{\xi_1, \dots, \xi_k\}\right) := \rho\left(\eta_1 - \hat{\eta}_1^*, \eta_2 - \hat{\eta}_2^*\right), \quad \text{где } \hat{\eta}_i^* = \underset{\hat{\eta}_i \in \left\{\sum_{i=1}^k b_i \xi_i + b_0\right\}}{\operatorname{argmin}} \mathsf{E}\left(\eta_i - \hat{\eta}_i\right)^2.$$

Если регрессия линейна, то

$$\rho(\eta_1, \eta_2 \mid \xi_1, \dots, \xi_k) = \rho(\eta_1 - \mathsf{E}\left\{\eta_1 \mid \xi_1, \dots, \xi_k\right\}, \eta_2 - \mathsf{E}\left\{\eta_2 \mid \xi_1, \dots, \xi_k\right\}).$$

Замечание (Важное). Пусть в эксперименте подсчитан ненулевой ρ. Это может означать, что один из факторов является причиной, а другой следствием; чтобы установить, что есть что, проводят эксперимент и смотрят, какой фактор в реальности влияет на какой. Это может также означать, что влияет сторонний фактор. Чтобы его исключить, считают частную корреляцию.

Пример. Возможна ситуация, когда $\rho(\eta_1, \eta_2) \neq 0$, но $\rho(\eta_1, \eta_2 \mid \xi) = 0$. Частная корреляция есть, по сути, корреляция на центрированных данных.

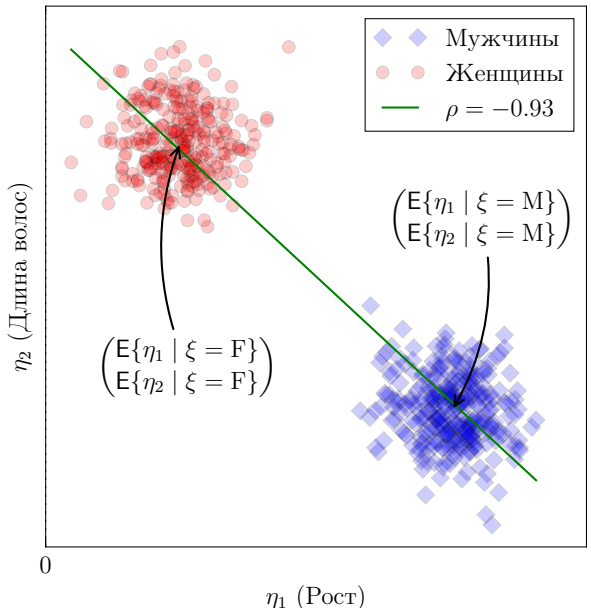

Рис. 5.1.: Исходные данные (бимодальность)

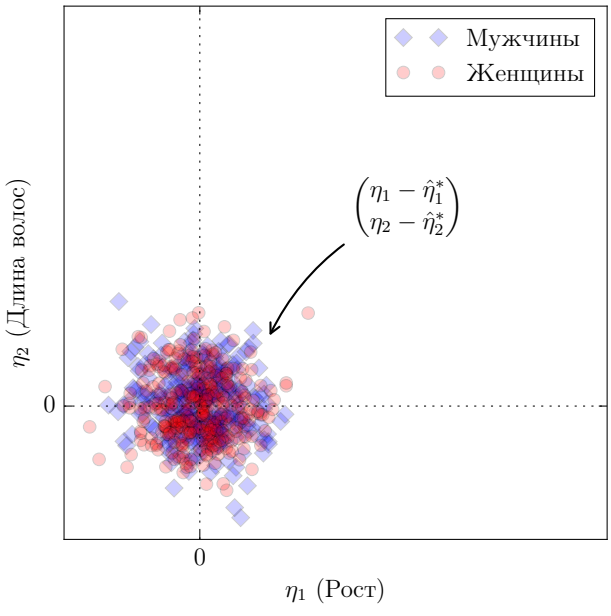

Рис. 5.2.: Центрированные данные

Пример. Возможна и ситуация как на (5.3), где определенно $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$.

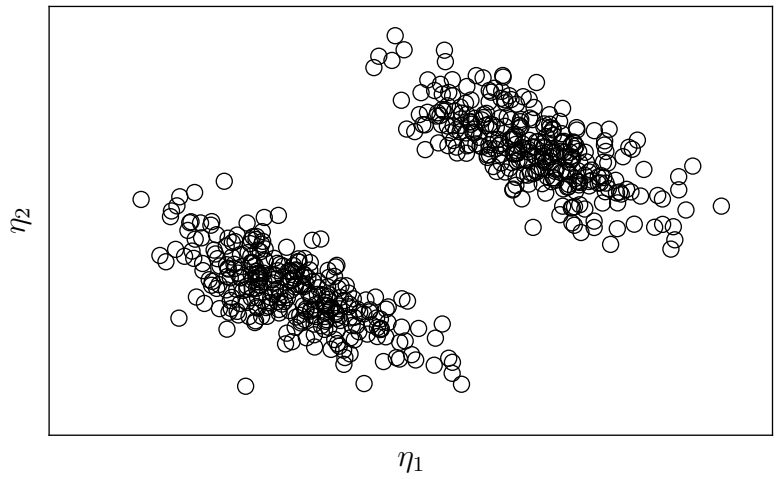

Рис. 5.3.: $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$

Замечание. По аналогии с предыдущим примером, если $|\operatorname{im} \xi| \to \infty$, то графики (η_1, η_2) при фиксированном ξ образуют эллипсоид (в этом случае с положительной корреляцией).

6. Зависимость между порядковыми признаками

Пусть на выборке

$$\binom{\xi}{\eta} \sim \binom{x_1}{y_1}, \dots, \binom{x_n}{y_n}$$

задан только порядок. Тогда можем считать только эмпирическую функцию распределения.

Следующие коэффициенты основаны на рангах. Ранговые характеристики хорошо работают на выборках *без повторений* (чтобы, к примеру, не возникало дробных рангов).

6.1. Ранговый коэффициент Спирмана

Определение. *Ранговый коэффициент Спирмана* есть

$$\rho_S = \rho(\mathrm{cdf}_\xi(\xi), \mathrm{cdf}_\eta(\eta)).$$

Замечание. $\mathrm{cdf}_\xi(\xi) \sim \mathrm{U}(0,1)$, потому что $\mathsf{P}(\mathrm{cdf}_\xi(\xi) < x) = \mathsf{P}(\xi < \mathrm{cdf}_\xi^{-1}(x)) = \mathrm{cdf}_\xi(\mathrm{cdf}_\xi^{-1}(x)) = x$.

Определение. *Ранг* элемента из выборки есть его порядковый номер в упорядоченной выборке:

$$\mathrm{rk}\, x_{(i)} = i.$$

Обозначение. $\mathrm{rk}\, x_{(i)} =: R_i$, $\mathrm{rk}\, y_{(i)} =: T_i$.

Можем ввести эмпирическое распределение

$$\mathrm{cdf}_{\xi_n}(x_i) = \frac{\mathrm{rk}\, x_i}{n}, \quad \mathrm{cdf}_{\eta_n}(y_i) = \frac{\mathrm{rk}\, y_i}{n} = \frac{T_i}{n}.$$

Тогда будет справедливо следующее

Определение. *Выборочный коэффициент Спирмана* определяется как выборочный коэффициент корреляции Пирсона $\hat\rho$, но с заменой значений на ранги:

$$\hat\rho_S \binom{\xi_n}{\eta_n} = \rho \binom{R_n}{T_n} = \frac{1/n \cdot \sum_{i=1}^n R_i T_i - \bar R \bar T}{\sqrt{1/n \cdot \sum_{i=1}^n \left(R_i - \bar R\right)^2} \sqrt{1/n \cdot \sum_{i=1}^n \left(T_i - \bar T\right)^2}}.$$

Если нет повторяющихся наблюдений, то знаменатель будет одним и тем же у всех выборок объема n, значит его можно посчитать заранее. В этом (и только этом) случае, справедлива более простая формула:

$$\hat\rho_S = 1 - \frac{6 \sum_{i=1}^n (R_i - T_i)^2}{n^3 - n}.$$

Замечание. Из последней формулы хорошо видно, что если x_i, y_i все идут в одном порядке, то $R_i - T_i = 0\ \forall i$ и $\hat\rho_S = 1$.

Замечание. ρ_S для *количественных* признаков есть мера монотонной зависимости:

$$\rho_S = 1 \iff (x_i > x_{i+1} \implies y_i > y_{i+1}\ \forall i)$$

(даже если зависимость нелинейная и $\rho \neq 1$). Иными словами, $\rho_S > 0$, если y имеет тенденцию к возрастанию с возрастанием x (и $\rho_S < 0$ иначе). Чем большое ρ_S, тем более явно выражена зависимость y от x в виде некоторой монотонной функции.

6.1.1. Согласованость ρ и ρ_S

ρ_S не согласована с ρ в том же смысле, что ρ и $r_{\xi|\eta}$.

Утверждение. Если данные нормальные то справедлива формула

$$\rho = 2\sin\left(\frac{\pi}{6}\rho_S\right).$$

Значит, можем сравнить критерии между собой.

- С точностью до погрешности, по значению, $\hat{\rho}$ и $\hat{\rho}_S$ — это одно и то же (см. 6.1)

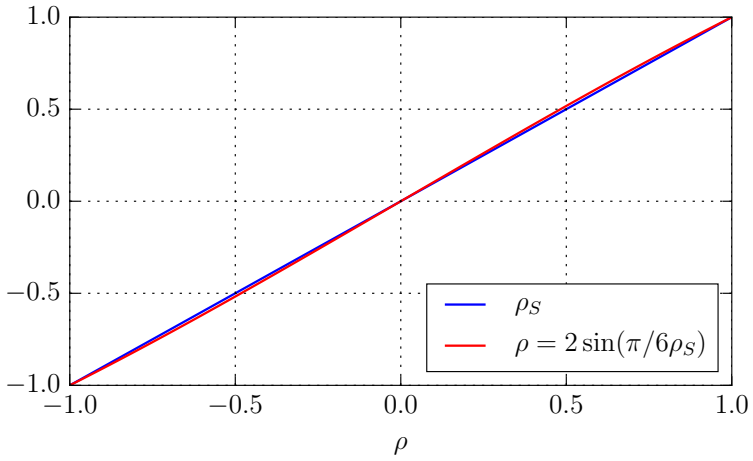

Рис. 6.1.: $\hat{\rho} \approx \hat{\rho}_S$

- Обычный критерий оценки — выборочную дисперсию — посчитать сложно. Тем не менее, можем заметить, что $\hat{\rho}_S$ более устойчив к аутлаерам (см. 6.2). Всегда можно добавить аутлаер такой, что $\hat{\rho} = 0$; $\hat{\rho}_S$ же поменяется не сильно. Поэтому для нормальных данных, ρ_S — это оценка, что нет аутлаеров.

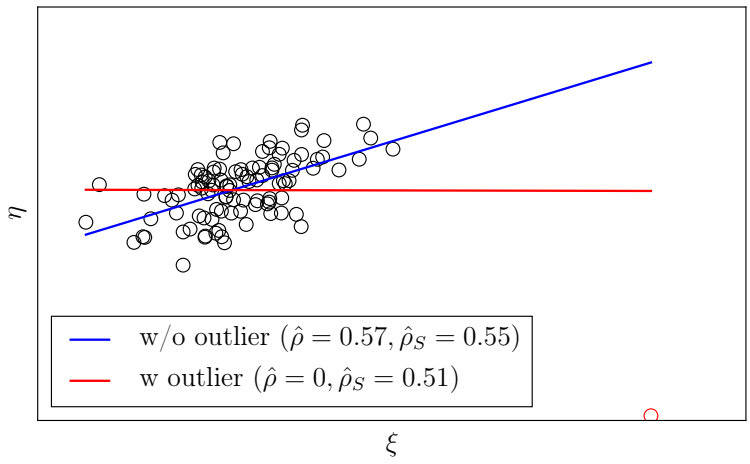

Рис. 6.2.: $\hat{\rho}$ до и после добавления аутлаера

- Монотонным преобразованием можем всегда сделать так, чтобы ρ изменился (например, возведя в квадрат); при монотонном преобразовании, однако, не меняется ρ_S (см. 6.3). Значит, чтобы узнать ρ исходных (нормальных) данных, можно не выполнять обратного преобразования, а сразу посчитать ρ_S.

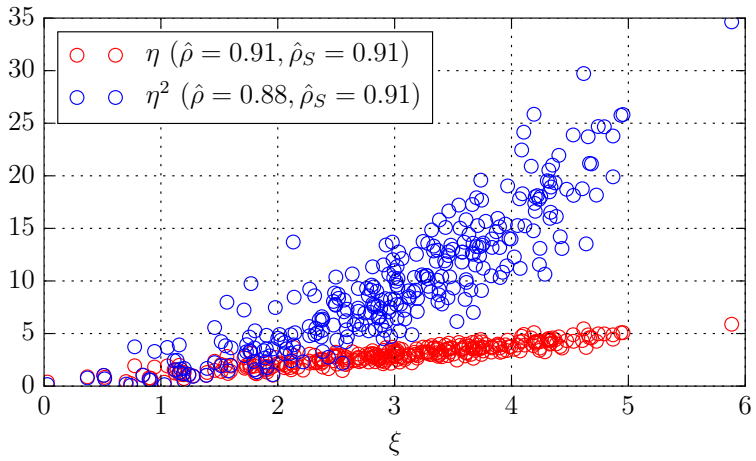

Рис. 6.3.: Монотонное преобразование нормальных данных

6.2. Ранговый коэффициент Кэндалла $\tau(\xi, \eta)$

Определение. Пусть $(\xi_1, \eta_1)^\mathsf{T} \perp\!\!\!\perp (\xi_2, \eta_2)^\mathsf{T} \sim \mathcal{P}_{\xi,\eta} \sim (\xi, \eta)^\mathsf{T}$; тогда *ранговым коэффициентом Кэндалла* называется

$$\tau(\xi, \eta) = \rho(\operatorname{sign}(\xi_2 - \xi_1), \operatorname{sign}(\eta_2 - \eta_1)) = \mathsf{P}((\xi_2 - \xi_1)(\eta_2 - \eta_1) > 0) - \mathsf{P}((\xi_2 - \xi_1)(\eta_2 - \eta_1) < 0).$$

На выборочном языке, пусть дана выборка $(x_1, y_1), \ldots, (x_n, y_n)$; тогда

$$\tau = \frac{\#(\text{одинаково упорядоченных пар}) - \#(\text{по-разному упорядоченных пар})}{\#(\text{комбинаций пар})},$$

где пара $(x_i, y_i), (x_j, y_j)$ считается одинаково упорядоченной, если $\operatorname{sign}(x_i - x_j) = \operatorname{sign}(y_i - y_j)$, а $\#(\text{комбинаций пар}) = C_n^2 = n(n-1)/2$.

Утверждение. Если $(\xi, \eta)^\mathsf{T} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, то справедлива формула

$$\rho = \sin\left(\frac{\pi}{2}\tau\right).$$

Из утверждения следует, что τ все время меньше ρ и ρ_S (по модулю).

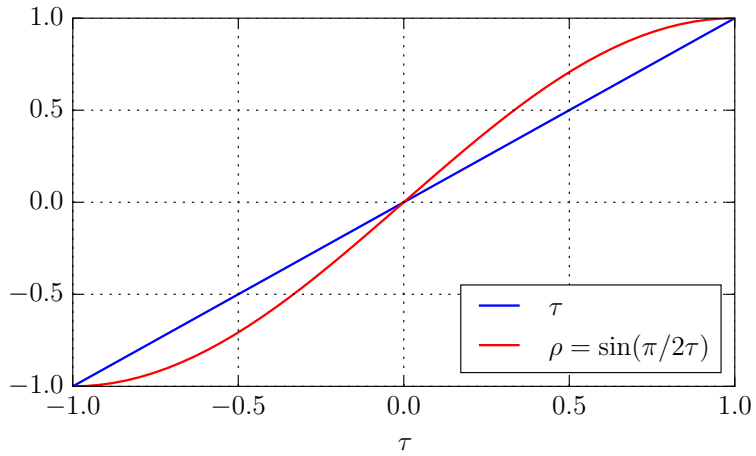

Рис. 6.4.: ρ и τ

Пример (Проверка ряда на тренд)**.** Пусть ξ — номера точек, а η — значения ряда. Тогда H_0 : $\tau_0 = 0$ и если H_0 отвергается, то тренд присутствует.

7. Корреляционные матрицы

Если признаков много, то их наглядно характеризуют корреляционные матрицы. Улучшить наглядность можно переупорядочив признаки так, чтобы на диагонали матрицы стояли блоки корреляций признаков из «корреляционных плеяд».

Определение. Пусть ρ_0; корреляционная плеяда есть множество признаков, таких, что их попарная корреляция больше ρ_0.

Можно выделить и несколько уровней ρ_i : $\rho_0 < \rho_1 < \dots$. Тогда сначала следует составить плеяду по ρ_0, затем внутри полученного по ρ_1 и т.д.

Часть V.

Дисперсионный анализ

1. Однофакторный дисперсионный анализ (One-way ANOVA[1])

Задача может быть поставлена двумя эквивалентными образами:

1. Пусть $\eta_i \sim \mathcal{P}_i,\ i \in 1:k$. Проверить гипотезу, что все распределения равны:
$$H_0 : \mathcal{P}_1 = \cdots = \mathcal{P}_k.$$
2. Пусть дан двумерный вектор $\begin{pmatrix}\xi & \eta\end{pmatrix}^\mathsf{T}$, причем ξ («фактор») принимает k значений $A_1, \ldots, A_k$. Рассмотрим $\eta_i \sim \mathcal{P}_i = \mathcal{P}_{\eta|\xi=A_i}$. Проверить гипотезу
$$H_0 : \mathcal{P}_{\eta|\xi=A_1} = \cdots = \mathcal{P}_{\eta|\xi=A_k}.$$

Пусть теперь $\eta_i \sim \mathrm{N}(\mu_i, \sigma^2)$. Разумеется,

$$H_0 : \mu_1 = \cdots = \mu_k \iff H_0 : \mathsf{E}\eta_1 = \cdots = \mathsf{E}\eta_k$$
$$\iff H_0 : \mathsf{E}(\eta \mid \xi = A_1) = \cdots = \mathsf{E}(\eta \mid \xi = A_k) \iff H_0 : \mathsf{DE}(\eta \mid \xi) = 0.$$

Для построения критерия, вспомним разложение дисперсии на выборочном языке:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{Q=\hat{\mathsf{D}}\eta} = \underbrace{\sum_{i=1}^{k} n_i\left(\bar{y}_i-\bar{y}\right)^2}_{Q_1=\hat{\mathsf{DE}}(\eta|\xi)} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}_{Q_2},$$

откуда в качестве критерия (следуя гипотезе) выберем Q_1 с идеальным значением 0. Однако Q_1 полезно отнормировать по Q_2 для учета различных внутригрупповых разбросов. Чтобы получить статистику с известным распределением, вспомним, что по теореме Cochran, $Q_1 \perp\!\!\!\perp Q_2$,

$$\frac{Q}{\sigma^2} \sim \chi^2(n-1), \quad \frac{Q_1}{\sigma^2} \sim \chi^2(k-1), \quad \frac{Q_2}{\sigma^2} \sim \chi^2(n-k)$$

и

$$t = \frac{Q_1/(k-1)}{Q_2/(n-k)} \sim \mathrm{F}((k-1),(n-k)).$$

Замечание. Это обобщение статистики для проверки гипотезы о равенстве математических ожиданий независимых двумерных выборок с равными дисперсиями (с $k = 2$, то есть):

$$t = \frac{\bar{x}-\bar{y}}{\tilde{s}_{1,2}\sqrt{1/n_1 + 1/n_2}}$$

с $\tilde{s}^2_{1,2} = Q_2/(n-2)$. Дело в том, что статистики распределены одинаковы — по определению,

$$t^2(n-2) = \mathrm{F}(1, n-2).$$

Чтобы воспользоваться полученным критерием, должно убедиться, что дисперсии одинаковые. Как и в случае $k = 2$, это можно проверить по тесту Левена, только многомерному, т.е. проверить равенство математических ожиданий $\mathsf{E}\,|\xi - \mathsf{E}\xi_i|\ \ \forall i \in 1:k$, $|y_{ij} - \bar{y}_i|$ — опять же, через саму ANOVA.

[1] ANalysis Of VAriation

Замечание. Если условия нормальности нарушаются, то критерий становится асимптотическим. Тогда вместо F следует использовать χ^2, так как $\mathrm{F}(k,m)/m \xrightarrow[m\to\infty]{} \chi^2(k)$.

Пример. Пусть дана выборка вида $\{(\xi = \text{пол}, \eta = \text{вес})\}$. Выдвинем H_0 : вес не зависит от пола. Очевидно, что ξ — категориальная случайная величина, а η — количественная. Значения ξ разобьют всю выборку на две (?) группы. Тогда проверка гипотезы сведется к проверке равенства распределений в двух группах, $\mathcal{P}_{\eta|\xi=s_1} = \mathcal{P}_{\eta|\xi=s_2}$. В предположении, что $(\eta \mid \xi = s_i) \sim \mathrm{N}(\mu_i, \sigma^2)$, равенство распределений будет следовать из равенства математических ожиданий.

2. Множественные сравнения

Пример. Проблема множественных сравнений возникает, например, в следующих ситуациях.

- Пусть одна группа испытуемых принимает лекарство, а вторая нет. По завершению эксперимента две группы сравниваются по m показателям. Однако чем больше показателей сравнивается, тем больше вероятность того, что *хотя бы по одному* показателю будет совпадение (в силу случайности).

- Испытывают $m = 100$ монет на честность сериями по $n = 10$ бросков: $\{(\xi_1^{(1)}, \dots, \xi_{10}^{(1)}), \dots, (\xi_1^{(100)}, \dots, \xi_{10}^{(100)})\}$, иными словами $\left\{\eta^{(1)}, \dots, \eta^{(100)}\right\}$, где $\eta^{(i)} \sim \mathrm{Bin}(10, p)$. Проверить m гипотез $H_0^{(i)} : \eta^{(i)} \sim \mathrm{Bin}(10, 1/2)$, $i \in 1 : m$. Зафиксируем $\alpha = 0.05$. Тогда, учитывая $\mathrm{pmf}_{\mathrm{Bin}(10,1/2)}(k) = C_{10}^k 2^{-10}$, $\mathsf{P}_{H_0^{(i)}}(\eta^{(i)} \geq 9) = 10 \cdot 2^{-10} + 1 \cdot 2^{-10} \approx 0.0107$, однако уже $\mathsf{P}_{H_0^{(i)}}(\eta^{(i)} \geq 8) \approx 0.0546$. Так что критерием наибольшей мощности будет $\eta^{(i)} \geq 9$:

$$\alpha_1 = \mathsf{P}_{H_0^{(i)}}(H_0^{(i)} \text{ отв}) = \mathsf{P}_{H_0^{(i)}}(\eta^{(i)} \geq 9) \approx 0.0107 \leq 0.05.$$

  Но использование того же критерия для множественных сравнений сильно завышает α_{I}:

$$\begin{aligned} \mathsf{P}\left(\bigvee_{i=1}^{100} H_0^{(i)} \text{ отв}\right) &= 1 - \mathsf{P}\left(\bigwedge_{i=1}^{100} H_0^{(i)} \text{ не отв}\right) = 1 - \left(1 - \mathsf{P}(H_0^{(i)} \text{ отв})\right)^{100} \\ &= 1 - (1 - 0.0107)^{100} \approx 0.6589. \end{aligned}$$

Пусть проверяются гипотезы $H_0^{(1)}, \dots, H_0^{(m)}$. Возможны такие ситуации:

	Retain H_0 (критерий не значим)	Reject H_0 (критерий значим)
True H_0	# True Negative	# False Discovery
False H_0	# False Negative	# True Discovery

Используя обозначения таблички,

$$\alpha_{\mathrm{I}} \approx \frac{\mathrm{FD}}{\mathrm{TN} + \mathrm{FD}}, \qquad \alpha_{\mathrm{II}} \approx \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TD}}.$$

Определение. Family-wise error rate (FWER):

$$\mathrm{FWER} = \mathsf{P}(\text{хотя бы один раз отвергнута верная гипотеза}) = \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right).$$

Иными словами, FWER — это ошибка первого рода для всей совокупности экспериментов.

Требуется контролировать FWER на предзаданном уровне α, т.е. чтобы $\mathrm{FWER} \sim \alpha$, где $\sim \in \{=, \leq, \to\}$. В *слабом* смысле это осуществляется, если $\mathrm{FWER} \sim \alpha$ только если *все* $H_0^{(i)}$, $i \in 1 : m$ верны. В *сильном* смысле контроль FWER на уровне α гарантируется для *любой* конфигурации верных и не верных $H_0^{(j)}$.

Определение. Пусть $T_0 := \{i : H^{(i)} \text{ верна}\}$. Тогда

$$\text{weak FWER}_T = \mathsf{P}(\text{хотя бы один раз отвергнута верная гипотеза, если верны } H^{(i)},\ i \in T)$$

т.е. если $T_0 = T$.

Определение.

$$\text{strong FWER} = \max_{T:T\subset\{1,\dots,m\}} \text{weak FWER}_T.$$

Обозначение. $\text{FWER}_T := \text{weak FWER}_T$.

Это осуществляется двумя процедурами:

- Single
- Stepdown

2.1. Single

Каждая $H^{(i)}$ проверяется отдельно с уровнем значимости α_1. Задача сводится к тому, чтобы найти такое α_1, что $\text{FWER} \le \alpha$ для какого-то нужного предзаданного α. Пусть $T = 1 : m$, т.е. будто все тесты верны; тогда

$$\text{FWER}_{\{1,\dots,m\}} = \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right) \le \sum_{i=1}^{m} \mathsf{P}(H^{(i)} \text{ отв}) = m\alpha_1 = \alpha \implies \alpha_1 := \frac{\alpha}{m}.$$

Замечание. Из-за неравенства тест консервативный, т.е. $\text{FWER} \ll \alpha$. Значит не максимально мощный.

$$\begin{aligned} \text{strong FWER} &= \max_{T\subset\{1,\dots,m\}} \mathsf{P}(H^{(i)} \text{ отвергается},\ i \notin T) \\ &\le \sum_{i\notin T} \mathsf{P}(H^{(i)} \text{ отвергается}) = |\{i : i \notin T\}|\,\alpha_1 = \alpha. \end{aligned}$$

Следствие. FWER *всегда хуже* strong FWER.

Определение. Поправка Бонферрони

$$\alpha_1 = \frac{\alpha}{m}.$$

Тест нужно проверять не с α_1, а с α/m. Так критерий будет консервативным (иначе — радикальным, что хуже).

Определение. Поправка Бонферрони для p-value:

$$p\text{-value} < \frac{\alpha}{m} \implies \text{отвергаем} \iff mp < \alpha \implies \text{отвергаем}.$$

2.2. Stepdown (Holm's algorithm)

Для увеличения мощности применяется «Holm's algorithm»:

1. считаются все p-value $p_1, \dots, p_m$,
2. упорядочиваются: $p_{(1)} \le \dots \le p_{(m)}$.
3. если $mp_{(1)} < \alpha$ то гипотеза отвергается, иначе и все последующие не отвергаются

4. в общем, если
$$p_{(j)} < \frac{\alpha}{m-j+1}$$
то гипотеза отвергается, иначе и все последующие не отвергаются.

Замечание. Сей тест более мощный, потому что не всегда происходит умножение на m.

Замечание. Процедуру сложно повторить, потому что при упорядочивании гипотезы могут перемешиваться.

Предложение. $\mathrm{FWER} \leq \alpha$.

Доказательство. Упорядочим p-value: $p_{(1)} \leq \cdots \leq p_{(j)} \leq \ldots \leq p_{(m)}$. Пусть $I = \left\{i : H_0^{(i)} \text{ верна}\right\}$, $m_0 = |I|$ — количество верных гипотез, $j = \min_{k \in I}$. m_0 должно «поместиться до конца»:
$$j \leq m - m_0 + 1 \implies \frac{\alpha}{m-j+1} \leq \frac{\alpha}{m_0}.$$

Значит
$$\begin{aligned} \mathrm{FWER}_I &\leq \mathsf{P}\left(p_{(j)} < \frac{\alpha}{m-j+1}\right) \leq \mathsf{P}\left(p_{(j)} < \frac{\alpha}{m_0}\right) \leq \mathsf{P}\left(\min_{i \in I} p_i < \frac{\alpha}{m_0}\right) \\ &= \mathsf{P}\left(\bigvee_{i \in I} p_i < \frac{\alpha}{m_0}\right) \leq \sum_{i \in I} \mathsf{P}\left(p_i < \frac{\alpha}{m_0}\right) = m_0 \frac{\alpha}{m_0} = \alpha \end{aligned}$$

□

2.2.1. Частный случай

Если все гипотезы и критерии независимы, то возможно точно посчитать FWER:
$$\begin{aligned} \mathrm{FWER}_{\{1,\ldots,m\}} &= \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right) = 1 - \mathsf{P}\left(\bigwedge_{i=1}^{m} H_0^{(i)} \text{ не отв}\right) \\ &= 1 - (1-\alpha_1)^m = \alpha \implies \alpha_1 = 1 - \sqrt[m]{1-\alpha} \end{aligned}$$

Определение. *Поправка Šidák*:
$$\alpha_1 = 1 - \sqrt[m]{1-\alpha}.$$

3. ANOVA Post-Hoc Comparison

В случае отвержения гипотезы ANOVA, можно провести дополнительное выборочное тестирование выделенных групп.

3.1. Least Significant Difference (LSD)

LSD test — это просто попарный t-test:

$$t = \frac{\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j}{\tilde{s}_{1,\dots,k}\sqrt{1/n_i + 1/n_j}} \sim \mathrm{t}(n-k),$$

где $\tilde{s}_{1,\dots,k}$ — это pooled по k группам standard deviation.

Замечание. Его стоит применять после множественного сравнения лишь к тем группам, важность которых была зафиксирована экспериментатором до проведения множественного сравнения.

Замечание. Критерий радикален. Значит, если он не нашел разницу, то и другие критерии тоже не найдут.

Замечание. Если групп немного, то можно применить поправку Бонферрони.

3.2. Распределение размаха

Сопоставим $\xi_1, \dots, \xi_n$ i.i.d. с $\mathrm{cdf}_{\xi_i}(x) = F(x)$ вариационный ряд $\xi_{(1)}, \dots, \xi_{(n)}$.

Определение. *Размах* есть случайная величина

$$w_n = \xi_{(n)} - \xi_{(1)}$$

с функцией распределения

$$\mathsf{P}(w_n < w) = n \int_{\mathbb{R}} (F(x+w) - F(x))^{n-1} \,\mathrm{d}F(x)$$

($w_n < w \implies w_i < w$, $\mathsf{P}(w_i < w) = F(x+w) - F(x)$ — $n-1$ штук таких, плюс перебор разных минимумов по $1:n$).

Замечание. В частном случае $F(x) = \mathrm{cdf}_{\mathrm{N}(0,\sigma^2)}(x)$, $\Phi(x) = \mathrm{cdf}_{\mathrm{N}(0,1)}(x)$ рассматривается *стандартизированный размах*

$$\mathsf{P}\left(\frac{w_n}{\sigma} < w\right) = n \int_{\mathbb{R}} (\Phi(x+w) - \Phi(x))^{n-1} \,\mathrm{d}\Phi(x).$$

Если σ неизвестна, то с подставленной оценкой $w/\tilde{s}$ называется стьюдентизированным размахом.

Утверждение. Пусть ℓ — некий параметр и — η^2 такая, что $\ell\eta^2/\sigma^2 \sim \chi^2(\ell)$; тогда

$$\frac{w_n}{\eta} \sim q(n, \ell),$$

где q — распределение стьюдентизированного размаха. Это распределение затабулировано.

Пример (Проверка выборки на outliers)**.** В нормальной модели, H_0 : нет outliers. Статистика

$$\frac{x_{(n)} - x_{(1)}}{\tilde{s}} \sim q(n, n-1)$$

потому что, естественно,

$$\frac{(n-1)\tilde{s}^2}{\sigma^2} \sim \chi^2(n-1).$$

Замечание. Полученный критерий не очень мощный — если H_0 отвергается, то есть аутлаеры присутствуют, то $x_{(n)} - x_{(1)}$ есть большая величина, но аналогично большой является и $\tilde{s}$, поэтому всё значение статистики вырастет незначительно по сравнению со случаем не-отвержения H_0, когда аутлаеров нет. Мощность же тем больше, чем больше (по модулю) значение статистики в случае, когда требуется отвержение H_0. Это видно из того, что $\beta = \mathsf{P}_{H_1}(T(\mathbf{x}) \in \mathscr{A}_{\text{крит}})$; но мощность, как площадь под графиком плотности H_1 на критическом луче (которые располагаются на хвостах плотности H_0), тем больше, чем дальше плотность H_1 от H_0, т.е. чем больше значения статистики T в ситуации отвержения H_0.

Выход заключется в построении более устойчивых оценок для σ^2 — например, на основе медианы и абсолютного отклонения.

3.3. Tukey's Honest Signifficat Difference (HSD) Test

Предположение. *Модель нормальная с дисперсией σ_0^2, и дизайн сбалансирован:* $\mathrm{N}(\mu_i, \sigma_0^2)$, $n_0 = n_i \ \forall i \in 1:k$.

По предложению 3.2,

$$t = \frac{\bar{\mathbf{y}}_{(k)} - \bar{\mathbf{y}}_{(1)}}{\sqrt{2\tilde{s}^2_{1,\dots,k}/n_0}} \sim q(k, n-k).$$

Тогда для проверки $H_0 : \mu_i = \mu_j$ используется HSD статистика

$$t_{ij} = \frac{|\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j|}{\tilde{s}_{1,\dots,k}\sqrt{2/n_0}},$$

а p-value считаются по $q(k, n-k)$ (таким образом, смотрят на каждую пару $(\bar{\mathbf{y}}_i, \bar{\mathbf{y}}_j)$ как на пару из размаха).

Предложение. *Это точный критерий.*

Доказательство. Действительно

$$\begin{aligned}\mathrm{FWER}_{\{1:m\}} &= \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right) = 1 - \mathsf{P}\left(\bigwedge_{i=1}^{m} H_0^{(i)} \text{ не отв}\right) = 1 - \mathsf{P}\left(t_{ij} < t_\alpha \ \forall i, j\right) \\ &= 1 - \mathsf{P}\left(\max_{i,j} t_{ij} < t_\alpha\right) = 1 - \mathsf{P}\left(t_{k1} < t_\alpha\right) = 1 - \mathsf{P}\left(t_{k1} < F^{-1}(1-\alpha)\right) = 1 - (1-\alpha) = \alpha.\end{aligned}$$

□

3.4. Другие критерии

Newman-Keals stepdown вариант HSD.

Tukey-Cramer HSD вариант Tukey для несбалансированного дизайна

Dunnett сравнивает все группы с контрольной

3.5. Scheffé's Method

ANOVA гипотезу $H_0 : \mu_1 = \cdots = \mu_k$ можно записать как

$$H_0 : \sum_{i=1}^{k} c_i \mu_i = 0, \quad \sum_{i=1}^{k} c_i = 0,$$

где $\{c_i\}_{i=1}^{k}$ — *«контраст»*.

Пример. Пусть две группы принимают k лекарств, в том числе — первым номером — плацебо. Сравнить все лекарства с плацебо *одним сравнением* можно сравнив с ним среднее арифметическое всех лекарств, для чего положить $c_1 = 1,\ c_2 = \ldots c_k = -1/(k-1)$.

Полученную сумму следует отнормировать и получить статистику

$$t = \frac{\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_i}{\sqrt{\mathsf{D}\left(\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_i\right)}} = \frac{\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_i}{\sigma \sqrt{\sum_{i=1}^{k} c_i^2 / n_i}} \sim \mathrm{N}(0,1).$$

При замене σ на $\tilde{s}$, получают, *как обычно*, $t \sim \mathrm{t}(n-k)$.

Пусть $c_1, \ldots, c_d,\ d \le k-1$ — наборы ортогональных контрастов. Тогда для любого вектора

$$t_j = \frac{\sum_{i=1}^{k} c_i^{(j)} \bar{\mathbf{y}}_i}{\sigma \sqrt{\sum_{i=1}^{k} (c_i^{(j)})^2 / n_i}}, \quad j \in 1:d.$$

Линейная комбинация нормальных векторов с ортогональными коэффициентами независима. Следовательно, можно использовать поправки Šidák.

Сколько бы ни захотелось проверить контрастов, хочется уверенности, что $\mathrm{FWER} \le \alpha$. Статистика

$$\frac{t^2}{k-1} \sim \mathrm{F}(k-1, n-k).$$

Замечание. В HSD можно каждую пару рассматривать как конкретный набор контрастов. Следовательно, метод Шеффе менее мощный по сравнению с HSD (поскольку проверяет все).

3.6. Сравнение мощностей

Статистики всех критериев можно свести к одной с разными критическими значениями. Для примера, пусть $k = 4, n = 20, \alpha = 0.05$; тогда

Критерий	*Критическое значение*
LSD	2.09
Dunnett	2.54
Bonferroni с 3-мя плановыми сравнениями	2.63
HSD	2.8
Bonferroni с $6 = C_4^2$ сравнениями	2.93
Scheffé	3.05

Чем больше критическое значение, тем ниже мощность, конечно.

Часть VI.

Регрессионный анализ

1. Регрессия

Определение. *Регрессией* η на ξ называется $\mathsf{E}\{\eta \mid \xi\}$.

Замечание. Таким образом осуществляется предсказание η по ξ с минимальной среднеквадратичной ошибкой.

Определение. *Функция регрессии* есть $f(x) = \mathsf{E}\{\eta \mid \xi = x\}$.

Замечание. f находится по МНК для $K = \{\psi(\xi) : \psi\text{—измеримая}\}$.

Виды регрессий

- Нелинейные и линейные ($K = \{a\xi + b\}$);
- Парные (предсказывая величину по одной случайной величине) и множественные (по многим).

2. Парная линейная регрессия

Определение. Пусть $\xi, \eta \in L^2$. *Парной линейной регрессией η по ξ* называется наилучшее среднеквадратичное приближение $h_{\beta_1^*,\beta_2^*}(\xi) = \beta_1^*\xi + \beta_2^*$ в классе линейных по ξ функций $K = \mathcal{L} = \{\beta_1\xi + \beta_2\}$. Иными словами,

$$h_{\beta_1^*,\beta_2^*}(\xi) = \operatorname*{argmin}_{\beta_1,\beta_2} \|\eta - h_{\beta_1,\beta_2}(\xi)\|^2 = \mathsf{E}\{\eta \mid h_{\beta_1,\beta_2}(\xi)\}\} = \operatorname*{argmin}_{\beta_1,\beta_2} \underbrace{\mathsf{E}(\eta - (\beta_1\xi + \beta_2))^2}_{\phi(\beta_1,\beta_2)} = \beta_1^*\xi + \beta_2^*.$$

Замечание. Найти минимум ϕ можно, как обычно, решив систему $\partial\phi/\partial\beta_i = 0$[1].

Утверждение. β_1^*, β_2^* таковы, что

$$\frac{h(\xi) - \mathsf{E}\eta}{\sqrt{\mathsf{D}\eta}} = \rho\frac{\xi - \mathsf{E}\xi}{\sqrt{\mathsf{D}\xi}}.$$

Это уравнение задает линию регрессии. Иными словами,

$$h(\xi) = \underbrace{\rho\frac{\sqrt{\mathsf{D}\eta}}{\sqrt{\mathsf{D}\xi}}}_{\beta_1^*}\xi + \underbrace{\mathsf{E}\eta - \rho\frac{\sqrt{\mathsf{D}\eta}}{\sqrt{\mathsf{D}\xi}}\mathsf{E}\xi}_{\beta_2^*}.$$

Отсюда можно получить соотношение между коэффициентом линейной регрессии $\beta_1^* = k$ (наклоном регрессионной прямой) и коэффициентом корреляции:

$$k = \rho\frac{\sigma_\eta}{\sigma_\xi}.$$

Замечание. Подстановкой проверятся, что

$$\phi(\beta_1^*, \beta_2^*) = \min_{\hat\eta \in K} \mathsf{E}\,(\eta - \hat\eta)^2 = \mathsf{D}\eta(1 - \rho^2),$$

откуда можно найти уже известное выражение для коэффициента корреляции Пирсона

$$\rho^2(\eta, \xi) = 1 - \frac{\phi(\beta_1^*, \beta_2^*)}{\mathsf{D}\eta} = 1 - \frac{\min_{\hat\eta \in H} \mathsf{E}\,(\eta - \hat\eta)^2}{\mathsf{D}\eta}, \quad \hat\eta := h(\xi).$$

Определение. Линейная регрессия *значима*, если $\beta_1^* \neq 0 \implies \rho \neq 0$. Значимость регрессии эквивалентна значимости предсказания по ней.

Определение. Величина *sum of squares residual* есть

$$\mathrm{SSR} = n \cdot \phi(\beta_1^*, \beta_2^*) = \sum_{i=1}^{n} (y_i - \hat y_i)^2, \quad \hat y_i = h_{\beta_1^*,\beta_2^*}(x_i).$$

[1] См. https://en.wikipedia.org/wiki/Simple_linear_regression

2.1. Модель линейной регрессии

Можно описать выборку как

$$y_i = \beta_1 x_i + \beta_2 + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2),\ \epsilon_i \perp\!\!\!\perp \epsilon_j.$$

σ^2 — мешающий параметр, который можно оценить через SSR/n. Но если $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, то

$$\hat{\sigma}^2 = \frac{\mathrm{SSR}}{n-2}$$

есть несмещенная оценка σ^2. Значит,

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2(n-2).$$

Замечание. МНК минимизирует разницу $y_i - \hat{y}_i$, что на графике соответствует вертикальным отрезкам, соединяющим y_i и $\hat{y}_i = h(x_i)$. Это не то же, что минимизация перпендикуляров от y_i на $h(x)$ — техники метода анализа главных компонент («PCA»).

2.2. Доверительные интервалы для β_1 и β_2

Как обычно, помимо точечной оценки $\hat{\beta}_1$ и $\hat{\beta}_2$, интересуемся диапазоном значений, которые может принимать оценка с заданной вероятностью. Примем предположение о несмещенности оценки, т.е. $\mathsf{E}\hat{\beta}_i = \beta_i$. Поскольку в модели $y_i = \beta_1 x_i + \beta_2 + \epsilon_i$ ошибка $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ есть случайная величина, оценки $\hat{\beta}_i$ — тоже становятся случайными величинами: $\hat{\beta}_i \sim \mathrm{N}(\beta_i, \mathsf{D}\hat{\beta}_i)$. В курсе регрессионного анализа доказывается[2], что

$$\mathsf{D}\hat{\beta}_1 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \mathsf{D}\hat{\beta}_2 = \frac{\sigma^2}{n}.$$

Кроме того,

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\mathsf{D}\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{n} s_x} = \frac{\sqrt{\frac{\mathrm{SSR}}{n-2}}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, \qquad \mathrm{SE}(\hat{\beta}_2) = \mathrm{SE}(\hat{\beta}_1) \cdot s_x$$

Предложение. *Статистика*

$$t = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathrm{t}(n-2).$$

Доказательство. Известно,

$$t \sim \mathrm{t}(m) \iff t = \frac{\xi}{\sqrt{\eta/m}}, \quad \xi \sim \mathrm{N}(0,1),\ \eta \sim \chi^2(m).$$

Ясно, что

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathrm{N}(0,1), \qquad \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2(n-2).$$

[2] См. https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares

Тогда

$$\frac{\left(\dfrac{\hat\beta_1-\beta_1}{\left(\dfrac{\sigma}{\sqrt{\sum_{i=1}^n(x_i-\bar x)^2}}\right)}\right)}{\left(\dfrac{\left(\dfrac{\sqrt{\sum_{i=1}^n(y_i-\hat y_i)^2}}{\sigma}\right)}{\sqrt{n-2}}\right)}=\frac{\dfrac{(\hat\beta_1-\beta_1)\sqrt{\sum_{i=1}^n(x_i-\bar x)^2}}{\sigma}}{\dfrac{\sqrt{\sum_{i=1}^n\epsilon_i^2}}{\sigma\sqrt{n-2}}}=\frac{\hat\beta_1-\beta_1}{\sqrt{\dfrac{\frac{1}{n-2}\sum_{i=1}^n\hat\epsilon_i^2}{\sum_{i=1}^n(x_i-\bar x)^2}}}\sim \mathrm{t}(n-2).$$

□

Используя статистику t, введем доверительные интервалы с $c_\gamma=\mathrm{cdf}^{-1}_{\mathrm{t}(n-2)}\left((1+\gamma)/2\right)$:

$$t\in(-c_\gamma,c_\gamma)\iff\frac{\hat\beta_1-\beta_1}{\mathrm{SE}(\hat\beta_1)}\in(-c_\gamma,c_\gamma)\iff\beta_1\in\left(\hat\beta_1-c_\gamma\,\mathrm{SE}(\hat\beta_1),\hat\beta_1+c_\gamma\,\mathrm{SE}(\hat\beta_1)\right).$$

Аналогично, для β_2:

$$\beta_2\in\left(\hat\beta_2-c_\gamma\,\mathrm{SE}(\hat\beta_2),\hat\beta_2+c_\gamma\,\mathrm{SE}(\hat\beta_2)\right).$$

Замечание. На картинке доверительные интервалы изображаются в виде «рукавов» вокруг графика линейной регрессии — т.е. область всевозможных положений прямой при варьировании β_1,β_2 в заданных интервалах.

Пример. Линейная регрессия как предсказательная модель может быть использована неправильно в следующих случаях:

- неправильная модель;
- применение к неоднородным данным (аутлаер или неоднородность);
- хотим построить предсказание в точке, далекой от данных (проблема — большая ошибка);
- не знаем какая модель там, где данных нет.

3. Множественная линейная регрессия

3.1. Псевдо-обратные матрицы

Определение. Матрица $\mathbf{A}^-$ называется *псевдо-обратной*, если

1. По аналогии с $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \implies \mathbf{A}\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}$ и $\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}$, выполняется

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}, \quad \mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-.$$

2. *(Псевдо-обратная по Муру-Пенроузу)* По аналогии с $\mathbf{A}^{-1} = \mathbf{A}^\mathsf{T} \implies (\mathbf{A}^{-1}\mathbf{A})^\mathsf{T} = \mathbf{A}^\mathsf{T}\left(\mathbf{A}^{-1}\right)^\mathsf{T} = \mathbf{A}^{-1}\mathbf{A}$, выполняется
$$\mathbf{A}^-\mathbf{A} = (\mathbf{A}^-\mathbf{A})^\mathsf{T}, \quad \mathbf{A}\mathbf{A}^- = (\mathbf{A}\mathbf{A}^-)^\mathsf{T}.$$

Свойства

1. Если столбцы $\mathbf{A}$ линейно-независимы, то существует $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$ и

$$\mathbf{A}^- = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}.$$

2. Пусть ищут решение $\mathbf{X}\mathbf{b} = \mathbf{y}$ относительно $\mathbf{b}$

   a) Если уравнение не имеет решений, то на $\mathbf{b} = \mathbf{X}^-\mathbf{y}$ достигается минимум невязки между левой и правой частями:

$$\mathbf{b}^* = \mathbf{X}^-\mathbf{y} = \underset{\mathbf{b}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2.$$

   b) Если решение не единственно, то $\mathbf{b} = \mathbf{X}^-\mathbf{y}$ есть решение с минимальной нормой.

3.2. Проекторы на подпространства

Пусть $\mathscr{L}_d \subset \mathbb{R}^m$ — линейное подпространство размерности d, натянутое на $\{\mathbf{p}_1, \ldots, \mathbf{p}_d\}$, $\mathbf{P} = [\mathbf{p}_1 : \ldots : \mathbf{p}_d]$. Тогда проектор на $\mathscr{L}_d$ будет задан как

$$\underset{\mathscr{L}_d}{\operatorname{proj}} = \mathbf{\Pi} = \mathbf{P}(\mathbf{P}^\mathsf{T}\mathbf{P})^{-1}\mathbf{P}^\mathsf{T} = \mathbf{P}\mathbf{P}^-.$$

Если $\{\mathbf{p}_i\}_{i=1}^d$ — ортонормированная система, то

$$\mathbf{\Pi} = \mathbf{P}\mathbf{I}\mathbf{P}^\mathsf{T} = \mathbf{P}\mathbf{P}^\mathsf{T}.$$

Кроме того,

$$\underset{\mathscr{L}_d^\perp}{\operatorname{proj}} = \underset{m\times m}{\mathbf{I}} - \mathbf{P}\mathbf{P}^\mathsf{T}.$$

(т.е., чтобы получить ортогональное пространство к проекции, нужно из исходного вектора вычесть проекцию).

Свойства

1. $\mathbf{\Pi\Pi} = \mathbf{\Pi}$
2. $(\mathbf{I} - \mathbf{\Pi})(\mathbf{I} - \mathbf{\Pi}) = \mathbf{I} - \mathbf{\Pi}$
3. $\mathbf{\Pi}^\mathsf{T} = (\mathbf{P}\mathbf{P}^\mathsf{T})^\mathsf{T} = \mathbf{\Pi}$.

3.3. Ordinary and Total Least Squares

Пусть

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{1n} & \dots & x_{nk} \end{pmatrix}$$

матрица данных с n индивидами [1] по столбцам, каждый из которых описывается k признаками;

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

вектор наблюдений[2];

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}$$

вектор неизвестных коэффициентов.

OLS Пусть $\mathbf{y} \in \mathbb{R}^n,\ \mathbf{X} \in \mathbb{R}^{n\times m},\ \mathbf{b} \in \mathbb{R}^m,\ \operatorname{rk}\mathbf{X} = m$. Пусть допускаются ошибки в наблюдениях такие, что $\mathsf{E}\epsilon_i = 0,\ \epsilon_i \perp\!\!\!\perp \epsilon_j,\ \mathsf{D}\epsilon_i = \sigma^2 \implies \operatorname{cov}\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$. Тогда в модели

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon},$$

найти

$$\hat{\mathbf{b}} = \operatorname*{argmin}_{\mathbf{b}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2 = \operatorname*{argmin}_{\tilde{\mathbf{y}}} \|\tilde{\mathbf{y}} - \mathbf{y}\|^2 = \mathbf{X}^-\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \quad \tilde{\mathbf{y}} := \mathbf{X}\mathbf{b}.$$

Откуда регрессией будет[3]

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \underbrace{\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}}_{\substack{\text{proj} \\ \text{colspace}\,\mathbf{X}}}\mathbf{y} = \mathbf{H}\mathbf{y}.$$

Можно посчитать остатки — разницу между наблюдениями и предсказанием по регрессии:

$$\text{residuals} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y}.$$

TLS Модель допускает ошибки $\boldsymbol{\Delta}$ также и в $\mathbf{X}$,

$$\mathbf{y} = (\mathbf{X} + \boldsymbol{\Delta})\,\mathbf{b} + \boldsymbol{\epsilon}$$

(где известны $\mathbf{y},\ \tilde{\mathbf{X}} := \mathbf{X} + \boldsymbol{\Delta}$, а $\mathbf{X}$ — нет). Найти

$$\operatorname*{argmin}_{\mathbf{b}:\tilde{\mathbf{y}}=\tilde{\mathbf{X}}\mathbf{b}} \left(\left\|\tilde{\mathbf{X}} - \mathbf{X}\right\|_F^2 + \|\tilde{\mathbf{y}} - \mathbf{y}\|^2 \right), \quad \|\mathbf{A}\|_F^2 = \sum_{i.j} a_{ij}^2.$$

Дальше рассматривается OLS.

[1] Также «predictors», «regressors», «controlled variables», «explanatory variables», «features», «inputs».
[2] Также «regressands», «response», «explaining variables», «outcome», «experimented variables».
[3] $\mathbf{H}$ — «hat matrix».

3.4. Свободный член

Видно, что $\mathbf{Xb} = \mathbf{y}$ задает СЛАУ, где каждое уравнение — прямая, проходящая через 0. Чтобы иметь возможность описывать случаи не-центрированных данных, пригодны два варинаты:

1. Ввести фиктивный столбец из единиц:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times m}, \quad m = k + 1.$$

2. Центрировать признаки.

Предложение. *Оба способы эквиваленты.*

Теорема (О делении регрессоров)**.** *Пусть* $\mathbf{X}$ *матрица данных с признаками («регрессорами») по столбцам,* $\mathbf{X} = [\mathbf{X}_1 : \mathbf{X}_2]$*,* $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2)^\mathsf{T}$*,* $\mathbf{M}_1 = \mathbf{I} - \mathbf{H}_1$*,* $\mathbf{H}_1 = \operatorname{proj}_{\operatorname{colspace} \mathbf{X}_1}$*. Тогда* $\hat{\mathbf{b}}_2$ *можно получить как регрессию* $\mathbf{M}_1\mathbf{y}$ *на* $\mathbf{M}_1\mathbf{X}_2$*. Остатки регрессии* $\mathbf{M}_1\mathbf{y}$ *будут такими же как остатки исходной.*

Доказательство. Без доказательства. □

Пусть $\mathbf{b} \in \mathbb{R}^m$, $\hat{\mathbf{b}} = \mathbf{X}^{-}\mathbf{y} = \left(\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k\right)$. Центрируем $\mathbf{X}$, вычитая среднее по каждому столбцу: $\mathbf{X}^{(c)} \in \mathbb{R}^{n \times k}$. Центрируем $\mathbf{y}$: $\mathbf{y}^{(c)} \in \mathbb{R}^n$; тогда $\hat{\mathbf{b}}^{(c)} = (\mathbf{X}^{(c)})^{-}\mathbf{y}^{(c)}$ и по теореме

$$\hat{\mathbf{b}}^{(c)} = \begin{pmatrix} \hat{b}_1^{(c)} \\ \vdots \\ \hat{b}_k^{(c)} \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_k \end{pmatrix}, \quad \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y}^{(c)} - \hat{\mathbf{y}}^{(c)}.$$

Следствие.

$$\hat{b}_0 = \bar{\mathbf{y}} - \sum_{i=1}^{k} \hat{b}_i \bar{x}_i.$$

3.5. Стандартизованные признаки

Если признаки изначально измерены в разных шкалах, то коэффициенты перед признаками можно интерпретировать как «важность».

Определение. Чтобы стандартизировать наблюдения, следует поделить центрированные столбцы на нормы каждого столбца, получится $\mathbf{X}^{(s)} \in \mathbb{R}^{n \times k}$; $\mathbf{y}^{(s)} = \mathbf{y}^{(c)} / \left\|\mathbf{y}^{(c)}\right\|$. Тогда

$$\hat{\mathbf{b}}^{(s)} = (\mathbf{X}^{(s)})^{-}\mathbf{y}^{(s)} = \left(\left(\mathbf{X}^{(s)}\right)^\mathsf{T} \mathbf{X}^{(s)}\right)^{-1} \left(\mathbf{X}^{(s)}\right)^\mathsf{T} \mathbf{y}^{(s)} = \hat{\boldsymbol{\beta}}, \quad \hat{\beta}_i = \frac{\left\|\mathbf{x}_i^{(c)}\right\|}{\left\|\mathbf{y}^{(c)}\right\|} \hat{b}_i.$$

Вектор $\hat{\boldsymbol{\beta}}$ имеет такой вид, потому что по ходу вычислений два раза поделили и один раз умножили на $\left\|\mathbf{x}_i^{(c)}\right\|$, и умножили на $\left\|\mathbf{y}^{(c)}\right\|$.

3.6. Свойства оценки $\hat{\mathbf{b}}$

1. *Несмещенность* (по $\mathsf{E}\boldsymbol{\epsilon} = \mathbf{0}$):

$$\mathsf{E}\hat{\mathbf{b}} = \mathsf{E}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathsf{E}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathsf{E}\left(\mathbf{Xb} + \boldsymbol{\epsilon}\right) = \mathbf{b}.$$

2. Ковариационная матрица:

$$\operatorname{cov}\hat{\mathbf{b}} = \operatorname{cov}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\operatorname{cov}\boldsymbol{\epsilon}\,\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}.$$

3. *Состоятельность*: если оценка несмещенная и состоятельная в среднеквадратичном смысле, то она несмещенная; однако ситуация

$$\operatorname{MSE}\hat{\mathbf{b}} = \mathsf{D}\hat{\mathbf{b}} \xrightarrow[n\to\infty]{} 0$$

невозможна в текущей постановке, потому что $\mathbf{X}$ — фиксированная матрица наблюдений.

Предложение (О состоятельности оценки)**.** *Пусть $\mathbf{X}_n \in \mathbb{R}^{n\times m}$ — последовательность (случайных) матриц, $\boldsymbol{\epsilon}_n \in \mathbb{R}^n$; кроме того,*

a) Выполняется сильная регулярность независимых переменных:

$$\frac{1}{n}\mathbf{X}_n^\mathsf{T}\mathbf{X}_n \xrightarrow{\mathsf{P}} \mathbf{A}, \quad \mathbf{A} \text{ невырожденная}$$

b) Ошибки независимы с регрессорами

$$\frac{1}{n}\mathbf{X}_n^\mathsf{T}\boldsymbol{\epsilon}_n \xrightarrow{\mathsf{P}} \mathbf{0}_m.$$

Тогда оценка

$$\hat{\mathbf{b}}_{\mathrm{OLS},n} \xrightarrow{\mathsf{P}} \mathbf{b}$$

является состоятельной.

Доказательство.

$$n\left(\mathbf{X}_n^\mathsf{T}\mathbf{X}_n\right)^{-1} = \mathbf{A}^{-1} \implies \mathsf{E}(\hat{\mathbf{b}}_n - \mathbf{b})(\hat{\mathbf{b}}_n - \mathbf{b}) = \operatorname{cov}\hat{\mathbf{b}}_n = \sigma^2(\mathbf{X}_n^\mathsf{T}\mathbf{X}_n)^{-1} \xrightarrow{\mathsf{P}} \mathbf{0}.$$

Значит, оценка состоятельна в среднеквадратичном, значит состоятельна. □

Предложение (Об асимптотической нормальности оценки)**.** *Пусть* $\{\epsilon_i\}$ i.i.d.,

$$\mathbf{A}_n = \frac{1}{\sigma}\left(\mathbf{X}_n^\mathsf{T}\mathbf{X}_n\right)^{-1/2}\mathbf{X}_n^\mathsf{T}.$$

$\hat{\mathbf{b}}_{\mathrm{OLS}}$ *асимптотически нормальна тогда и только тогда, когда*

$$\max_i\left\{\mathbf{A}_{ni1}^2, \ldots, \mathbf{A}_{nin}^2\right\} \xrightarrow[n\to\infty]{} 0$$

3.7. Свойства $\hat{\mathbf{b}}^{(c)}$ и $\hat{\mathbf{b}}^{(s)}$

1. $\left(\mathbf{X}^{(c)}\right)^\mathsf{T}\mathbf{X}^{(c)}/n = \mathbf{S}_{\mathbf{xx}}$, $\left(\mathbf{X}^{(c)}\right)^\mathsf{T}\mathbf{y}^{(c)}/n = \mathbf{S}_{\mathbf{xy}}$ суть выборочные ковариационные матрицы (это можно вручную расписать и убедиться); тогда в их терминах

$$\hat{\mathbf{b}}^{(c)} = (\hat{b}_1, \ldots, \hat{b}_k)^\mathsf{T} = \left(\left(\mathbf{X}^{(c)}\right)^\mathsf{T}\mathbf{X}^{(c)}/n\right)^{-1}\left(\mathbf{X}^{(c)}\right)^\mathsf{T}\mathbf{y}^{(c)}/n = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}.$$

Следствие. *Чем более скоррелированы признаки, тем более пропорциональны столбцы $\mathbf{S}_{\mathbf{xx}}$ и тем более вырождена $\mathbf{S}_{\mathbf{xx}}$, значит «больше» $\mathbf{S}_{\mathbf{xx}}^{-1}$, следовательно $\hat{\mathbf{b}}^{(c)}$ и разница $\mathbf{y}^{(c)} - \hat{\mathbf{y}}^{(c)} = \mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\mathbf{b}}^{(c)}$.*

2. Ковариационная матрица:

$$\operatorname{cov}\hat{\mathbf{b}}^{(c)} = \sigma^2\big(\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)}\big)^{-1} = \frac{\sigma^2}{n}\cdot\mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow[n\to\infty]{} 0$$

3. Аналогично,

$$\hat{\mathbf{b}}^{(s)} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{R}_{\mathbf{xy}}$$

и

$$\operatorname{cov}\hat{\mathbf{b}}^{(s)} = \frac{\sigma^{(s)2}}{n}\mathbf{R}_{\mathbf{xx}}^{-1}, \quad \sigma^{(s)} = \frac{\sigma}{\left\|\mathbf{y}^{(c)}\right\|}.$$

3.8. Сравнение оценок

По аналогии с одномерным случаем, *наилучшая оценка* — с минимально возможной дисперсией; аналог дисперсии — ковариационная матрица. Порядок вводится следующим образом:

Определение. $\mathbf{A} < \mathbf{B} \iff \mathbf{A} - \mathbf{B}$ отрицательно определена, т.е.

$$\forall\boldsymbol{\gamma}\ \boldsymbol{\gamma}^{\mathsf{T}}(\mathbf{A}-\mathbf{B})\boldsymbol{\gamma} < 0.$$

Замечание. Пусть $\boldsymbol{\gamma}^{(i)} = (0,\dots,\underbrace{1}_{i},\dots,0)^{\mathsf{T}}$; тогда $a_{ii} < b_{ii}$.

Теорема (Гаусс-Марков)**.** *В условиях* $\mathsf{E}\epsilon_i = 0,\ \mathsf{D}\epsilon_i = \sigma^2,\ \epsilon_i \perp\!\!\!\perp \epsilon_j$, $\hat{\mathbf{b}}_{\mathrm{OLS}}$ *является «BLUE»: «best linear unbiased estimate».*

Следствие. *Если* $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$*, то*

$$\hat{\mathbf{b}}_{\mathrm{OLS}} = \hat{\mathbf{b}}_{\mathrm{MLE}}.$$

Доказательство. MLE оценка есть

$$\begin{aligned}\mathsf{P}(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} = \sigma^2\mathbf{I}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det\sigma^2\mathbf{I}}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})\,\sigma^{-2}\mathbf{I}(\mathbf{y}-\boldsymbol{\mu})^{\mathsf{T}}\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{X}\mathbf{b}-\boldsymbol{\mu}\|^2\right\} \underset{\mathbf{b}}{\to} \max\end{aligned}$$

что аналогично,

$$\|\mathbf{X}\mathbf{b}-\boldsymbol{\mu}\|^2 \underset{\mathbf{b}}{\to} \min.$$

Это же есть постановка задачи OLS. □

3.9. Разложение суммы квадратов и оценка σ^2

Обозначим SSE = SSError = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Пусть ошибки имеют нормальное распределение $N(0,\sigma^2)$. Тогда, с помощью теореме Cochran можно получить (без док-ва):

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(\underbrace{n-m}_{n-k-1})$$

и оценкой методом подстановки для σ^2 будет SSE/n; несмещенной оценкой (с поправкой на число степеней свободы) будет

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-m}.$$

3.10. Проверка значимости коэффициентов линейной регрессии и доверительных интервалов

Определение. Коэффициент b_i *значим*, если отвергается $H_0 : b_i = 0$. Если коэффициент значим, значит признак существенен для регрессии.

Для построения точного критерия, предполагают $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Значит, поскольку $\hat{\mathbf{b}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})$, $\hat{\mathbf{b}}$ имеет тоже нормальное распределение со средним $\mathbf{0}$ (по несмещенности), но какой-то ковариационной матрицей: $\hat{\mathbf{b}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Тогда $\mathsf{E}\hat{b}_i = b_i = 0,\ \mathsf{D}\hat{b}_i = \sigma_i^2$ и

$$t = \frac{\hat{b}_i - b_i}{\sqrt{\mathsf{D}\hat{b}_i}} = \frac{\hat{b}_i}{\sqrt{\sigma^2((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{ii}}} = \frac{\hat{b}_i}{\sqrt{\sigma^2/n \cdot \left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}}} = \sqrt{n}\frac{\hat{b}_i}{\sigma\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}} \sim \mathrm{N}(0,1).$$

Заметим, что в $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ нумерация идет от 0, а в $\mathbf{S}_\mathbf{x}$ от 1. Подставляя оценку σ, получают

$$t = \sqrt{n}\frac{\hat{b}_i}{\hat{\sigma}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}} = \sqrt{n}\frac{\hat{b}_i}{\sqrt{\dfrac{\mathrm{SSE}}{(n-m)}}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}} = \frac{\dfrac{\sqrt{n}\hat{b}_i}{\sigma\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}}}{\sqrt{\dfrac{\mathrm{SSE}}{(n-m)\,\sigma^2}}} = \frac{\mathrm{N}(0,1)}{\sqrt{\dfrac{\chi^2(n-m)}{n-m}}} \sim \mathrm{t}(n-m).$$

3.10.1. Расстояние Махаланобиса

Если на прямой разброс удобно измерять стандартных отклонениях σ, то в многомерном пространстве аналогом такой характеристики является расстояние Махаланобиса (Mahalanobis distance).

Определение. Пусть $\mathbf{V}$ — неотрицательно определенная симметричная матрица; тогда *расстояние Махаланобиса* есть

$$r_M^2(\mathbf{x}, \mathbf{y}; \mathbf{V}) = (\mathbf{x}-\mathbf{y})^\mathsf{T}\mathbf{V}^{-1}(\mathbf{x}-\mathbf{y}).$$

Замечание. Если $\boldsymbol{\xi} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, то

$$\mathrm{pdf}_{\boldsymbol{\xi}}(\mathbf{x}) = C \cdot \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} = C \cdot \exp\left\{-\frac{1}{2}r_M^2(\mathbf{x}, \boldsymbol{\mu}; \mathbf{V})\right\}.$$

Для любых двух $\mathbf{x}_1, \mathbf{x}_2$ на линии уровня, $\mathrm{pdf}_{\boldsymbol{\xi}}(\mathbf{x}_1) = \mathrm{pdf}_{\boldsymbol{\xi}}(\mathbf{x}_2)$. Значит, $r_M^2(\mathbf{x}_1, \boldsymbol{\mu}; \mathbf{V}) = r_M^2(\mathbf{x}_2, \boldsymbol{\mu}; \mathbf{V})$, в то время, как Евклидово расстояние не обязано быть одинаковым из-за разной выраженности главных компонент. Однако $r_M^2(\mathbf{x}, \mathbf{y}; \mathbf{I}) = \|\mathbf{x}-\mathbf{y}\|_2^2$. Таким образом, r_M^2 — это Еквлидово расстояние с поправкой на ковариацию, задаваемую $\mathbf{V}$.

Предложение. *Если* $\boldsymbol{\xi} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, *то*

$$r_M^2(\boldsymbol{\xi}, \boldsymbol{\mu}; \mathbf{V}) = (\boldsymbol{\xi}-\boldsymbol{\mu})^\mathsf{T}\mathbf{V}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu}) \sim \chi^2(m)$$

как сумма квадратов центрированных и нормированных нормальных случайных величин.

Действительно,

$$\boldsymbol{\eta} = \mathbf{V}^{-1/2}(\boldsymbol{\xi}-\boldsymbol{\mu}) \sim \mathrm{N}(0, \mathbf{I}) \implies r_M^2(\boldsymbol{\xi}, \boldsymbol{\mu}; \mathbf{V}) = r_M^2(\boldsymbol{\eta}, \mathbf{0}; \mathbf{I}) = \boldsymbol{\eta}^\mathsf{T}\boldsymbol{\eta} \sim \chi^2(m).$$

3.10.2. Доверительный эллипсоид

В одномерном случае симметричного распределения, область носителя, где лежит γ всех значений распределения определяется равенством

$$\mathsf{P}(|\xi - \mathsf{E}\xi| < \sqrt{\mathsf{D}\xi}c_\gamma) = \gamma.$$

Т.е. как такое множество значений, что расстояние их от среднего с учетом стандартного отклонения меньше квантиля уровня γ. В случае оценки среднего μ_0, например, получают стандартное с $\mathrm{SE} = \sqrt{\mathsf{D}\bar{x}}$

$$\mathsf{P}\left(\frac{|\bar{x} - \mu_0|}{\mathrm{SE}} < c_\gamma\right) = \mathsf{P}\left(-c_\gamma < \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} < c_\gamma\right), \quad \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} \sim \mathrm{N}(0,1)$$

так что $c_\gamma = \mathrm{qnt}_{\mathrm{N}(0,1)}\,\gamma$.

Аналогично можно нарисовать m-мерный эллипсоид, в который помещается выборка с точностью γ. Расстояние с учетом ковариации будет задаваться соответственно параметризованным расстоянием Махаланобиса:

$$\mathsf{P}(r_M^2(\boldsymbol{\xi}, \boldsymbol{\mu}; \mathrm{SD}) < c_\gamma) = \gamma.$$

В случае, если $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, по предыдущему,

$$r_M^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) \sim \chi^2(m).$$

Значит,

$$\mathsf{P}(r_M^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) < c_\gamma) = \gamma, \quad c_\gamma = \mathrm{qnt}_{\chi^2(m)}\,\gamma.$$

3.11. Значимость регрессии

$H_0 : \mathbf{b}^{(c)} = \mathbf{0}$. Напомним что $\mathbf{b}^{(c)} = (b_1, \ldots, b_k)^{\mathrm{T}}$ и равенство его нулю означает то, что предсказание равно константе и не зависит от значений иксов. Эту гипотезу можно проверить тремя способами:

1. Аналогично парной регрессии: критерий

   $$t = r_M^2\left(\hat{\mathbf{b}}^{(c)}, \mathbf{0}; \mathrm{cov}(\hat{\mathbf{b}}^{(c)})\right) \sim \chi^2(k)$$

   а именно,

   $$t = \left(\hat{\mathbf{b}}^{(c)}\right)^{\mathsf{T}} \left(\mathrm{cov}(\hat{\mathbf{b}}^{(c)})\right)^{-1} \hat{\mathbf{b}}^{(c)} = \left(\hat{\mathbf{b}}^{(c)}\right)^{\mathsf{T}} \left(\frac{\sigma^2}{n} \cdot \mathbf{S}_{\mathbf{xx}}^{-1}\right)^{-1} \hat{\mathbf{b}}^{(c)} = \frac{n\left(\hat{\mathbf{b}}^{(c)}\right)^{\mathsf{T}} \mathbf{S}_{\mathbf{xx}} \hat{\mathbf{b}}^{(c)}}{\sigma^2}.$$

   Неизвестный σ^2 следует оценить как

   $$s^2 = \frac{\mathrm{SSE}}{n - (k+1)};$$

   тогда

   $$\frac{n\left(\hat{\mathbf{b}}^{(c)}\right)^{\mathsf{T}} \mathbf{S}_{\mathbf{xx}} \hat{\mathbf{b}}^{(c)}/k}{s^2} \sim \mathrm{F}(k, n - (k+1)).$$

2. Через ANOVA (разложение дисперсии): Разложение дисперсии

   $$\mathsf{D}\eta = \mathsf{E}(\eta - \mathsf{E}\eta)^2 = \mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2 + \mathsf{E}(\eta - \hat{\eta}^*)^2,$$

   где $\hat{\eta}^*$ — наилучшее линейное приближение от $\boldsymbol{\xi}$, на выборочном языке будет иметь вид

   $$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\mathrm{SSTotal}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\mathrm{SSRegr}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\mathrm{SSError}}.$$

   В случае, когда регрессоры не случайны (есть неслучайная матрица данных $\mathbf{X}$ и случайный отклик $\mathbf{y}$, как у нас сейчас), то же самое разложение имеет место.

Замечание. Иногда также пишут

$$\text{SSTotal} = \text{SSEffect} + \text{SSResidual},$$

что ведет к неиллюзорной путанице!

Пусть $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Тогда, с помощью теореме Cochran можно получить (без док-ва):

$$\frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1), \quad \frac{\text{SSR}}{\sigma^2} \sim \chi^2(\underbrace{m-1}_{k}), \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2(\underbrace{n-m}_{n-k-1})$$

и SSE $\perp\!\!\!\perp$ SSR.

Замечание. Утверждение про распределение SSE справедливо всегда при нормальном распределении ошибок; про SST и SSR это верно только если $\mathbf{b}^{(c)} = \mathbf{0}$. Именно поэтому применяется F-критерий для проверки значимости регрессии.

Таким образом, в качестве статистики F-критерия можно взять, как и в дисперсионном анализе,

$$t = \frac{\text{SSR}/k}{\text{SSE}/(n-(k+1))} \sim \mathrm{F}(k, n-(k+1))$$

Критическая область, очевидно, справа, так как 'идеальное значение' — 0.

Замечание. У этой статистики с предыдущей совпадает также и числитель, хотя, чтобы в этом убедиться, надо провести некоторые выкладки, так это не очевидно.

3. Через коэффициент детерминации регрессии: известно выражение для множественного коэффициента корреляции:

$$R^2(\eta; \xi_1, \ldots, \xi_k) = \frac{\mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2}{\mathsf{D}\eta}, \quad \mathsf{D}\eta = \mathsf{E}(\eta - \mathsf{E}\eta)^2 = \mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2 + \mathsf{E}(\eta - \hat{\eta}^*)^2;$$

на выборочном языке для множественной линейной регрессии получают

$$R^2 \quad = \quad \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}. \tag{3.1}$$

Если матрица регрессоров $\mathbf{X}$ фиксирована (т.е. не является выборкой из распределения $\boldsymbol{\xi}$), то R^2, вычисленный по той же формуле (3.1), называется коэффициентом детерминации.

Замечание. При удалении даже незначимого признака R^2 уменьшится; однако adjusted R^2

$$\text{adjusted}\, R^2 \quad = \quad 1 - \frac{\text{SSE}/(n-(k+1))}{\text{SST}/(n-1)} \xrightarrow[n\to\infty]{} R^2$$

не обязательно в силу поправки $n-(k+1)$, действующей как штраф за количество переменных.

Несложные манипуляции позволяют выписать статистику критерия ANOVA через коэффициент детерминации:

$$t = \frac{\dfrac{\text{SSR}}{k}}{\dfrac{\text{SSE}}{n-(k+1)}} = \frac{\dfrac{\text{SSR}}{k\,\text{SST}}\text{SST}}{\dfrac{(\text{SST}-\text{SSE}+\text{SST})\text{SST}}{\text{SST}}} = \frac{R^2/k}{\left(1-R^2\right)/(n-(k+1))}.$$

3.12. Анализ оценок коэффициентов

Для анализа оценок коэффициентов можно посмотреть на попарные срезы доверительного эллипсоида; точнее, на двумерные эллипсоиды. Для пары коэффициентов β_i, β_j его можно нарисовать (самостоятельно), в качестве центра взяв точку $(\hat\beta_i, \hat\beta_j)^\mathsf{T}$, наклон главной оси и вытянутость определив по величине $\operatorname{corr}(\hat\beta_i, \hat\beta_j)$.

- Чем дальше от начала координат эллипсоид, тем больше значимость признаков.
- Чем больше корреляция тем менее адекватно центр отражает ситуацию.
- Возможны два случая: когда эллипсоид перпендикулярен или сонаправлен прямым $y = \pm x$; в первом случае («хорошем») коэффициенты значимы в совокупности (даже если один близок к 0, то второй вполне далек и наоборот), во втором случае эллипсоид может довольно близко подходить к точек (0,0), т.е. оба коэффициента могут быть как одновременно малыми, так и большими (и, значит, и сильно, и слабо влиять на результат).

3.12.1. Корреляция между оценками коэффициентов в двумерном случае

При возрастании корреляции признаков:

- дисперсия оценок коэффициентов стремится к бесконечности;
- становится сложно оценить вклад каждого признака в регрессию.

Пример. Пусть $k = 2$, $\eta = b_0 + b_1\xi_1 + b_2\xi_2$. Пусть также матрица корреляций есть

$$\mathbf{R_{xx}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Тогда

$$\operatorname{cov}\hat{\boldsymbol\beta} = \frac{\sigma^{(s)2}}{n}\mathbf{R}_{\mathbf{xx}}^{-1} = \frac{\sigma^{(s)2}}{n}\frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

Значит, $\mathsf{D}\hat\beta_i \xrightarrow[\rho\to 1]{} \infty$.

3.12.2. Избыточность (redundancy) и ручное удаление признаков

С этой проблемой можно бороться, удаляя подходящие признаки из анализа[4] по следующим критериям:

1. Множественный коэффициент корреляции

   $$R^2(\xi_i; \{\xi_j,\ j \neq i\}).$$

   Чем он больше, тем скорее i-й признак нужно удалить.

2. Допустимость i-го признака:

   $$\text{tolerance}_i = 1 - R^2(\xi_i; \{\xi_j,\ j \neq i\}).$$

   Чем он меньше, тем скорее i-й признак нужно удалить. Помимо предыдущего соотношения справедливо

   $$\mathsf{D}\hat b_i = \frac{\sigma^2}{\sum_{\ell=1}^n (x_\ell - \bar{\mathbf{x}}_i)^2}\frac{1}{\text{tolerance}_i}, \quad \frac{1}{\text{tolerance}_i} \text{ — Variance Inflation Factor,}$$

   так что при маленькой допустимости дисперсия велика.

[4] Нет признака — нет проблемы.

3. Частные корреляции
$$\rho(\xi_i, \eta \mid \{\xi_j,\ j \neq i\}) = \rho(\xi_i - \hat{\xi}_i, \eta - \hat{\eta})$$
Чем i-я частная корреляция больше, тем больше вклад признака в регрессию (тем менее он предпочтителен для удаления).

4. Полу-частные корреляции
$$\rho(\xi_i - \hat{\xi}_i, \eta).$$

3.12.3. Проверка гипотезы о том, что набор признаков избыточен

Пусть $\mathbf{b} = (b_0, \ldots, b_{k-r}, \underbrace{b_{k-r+1}, \ldots, b_k}_{r \text{ штук}})^\mathsf{T}$. Если $H_0 : \mathbf{b}_{k-r+1,k} = \mathbf{0}$ не отвергается, значит последние r признаков не влияют на модель и следует выбрать более простую модель — без этих коэффициентов. Можно использовать расстояние Махаланобиса до 0 в метрике $\operatorname{cov}(\mathbf{b}_{k-r+1,k})$:

$$\begin{aligned} t &= r_M^2(\hat{\mathbf{b}}_{k-r+1,k}, \mathbf{0}; \operatorname{cov}(\mathbf{b}_{k-r+1,k})) \sim \chi^2(r) \\ &= \hat{\mathbf{b}}_{k-r+1,k}^\mathsf{T}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{(\mathrm{IV})}\hat{\mathbf{b}}_{k-r+1,k}/\sigma^2, \end{aligned}$$

где $((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{(\mathrm{IV})}$ — IV квадрант $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$. Если σ^2 неизвестна, то

$$\begin{aligned} t &= \frac{\hat{\mathbf{b}}_{k-r+1,k}^\mathsf{T}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{(\mathrm{IV})}\hat{\mathbf{b}}_{k-r+1,k}/r}{\hat{\sigma}^2} \sim \mathrm{F}(r, n-(k+1)) \\ &= \frac{(R^2_{1,k} - R^2_{1,k-r})/r}{(1-R^2_{1,k})/(n-m)}. \end{aligned}$$

3.12.4. Stepwise автоматическое удаление признаков

Выбор оптимального набора признаков можно производить автоматически, по одному добавляя признаки («Forward stepwise regression») или убирая их («Backward»). Пусть вариант Forward. На шаге i добавляется тот признак, что максимизирует

$$R^2_{1,i+1} - R^2_{1,i};$$

остановиться следует, когда $\left|R^2_{1,i+1} - R^2_{1,i}\right|$ достаточно мало. $H_0 : R^2_{1,i+1} - R^2_{1,i} = 0$, т.е. $b_{i+1} = 0$ перед добавленным признаком.

$$t = \frac{\hat{b}_i}{\mathrm{SE}(\hat{b}_i)} \sim \mathrm{t}(n-m).$$

Тогда $k = i+1$, $r = 1$ и статистика будет иметь вид

$$t^2 = \frac{(R^2_{1,i+1} - R^2_{1,i})}{(1-R^2_{1,i+1})/(n-(i+2))} \sim \mathrm{F}(1, n-(i+2)).$$

По сути, это есть перемасштабированное значение разницы $R^2_{1,i+1} - R^2_{1,i}$.

Замечание. Однако признак выбран «лучший» (а не случайный), значит распределение не F.

- Полное решение задачи — выбрать ℓ признаков из k перебором.
- Жадный алгоритм — последовательно выбирать наиболее подходящие признаки.

В Statistica есть критерий автоматической остановке для stepwise отбора признаков. F to enter в forward варианте — это пороговое значение для F, если F < этого числа, то останов. В backward варианте есть F to remove: если F > F to remove, то STOP. Только F to remove должно быть больше F to enter (на самом деле, на каждом шаге проверяется, можно ли добавить признак, а

потом какой-то другой удалить и неравенство нужно, чтобы процедура не зациклилась.) Имеет смысл установить такие пороги, чтобы критерий остановки не сработал, а потом посмотреть на таблицу Stepwise summary.

Можно нарисовать, как ведет себя коэффициент детерминации в варианте forward и backward. Это монотонные функции, но необязательно вторая производная одного знака. Отсюда можно увидеть, что критерий остановке в варианте backward более безопасен.

3.12.5. Выбор модели на основе информационных критериев AIC и BIC

См. отдельный файл.

3.12.6. О множественном коэффициенте корреляции и саппрессорах

Известно, что $\rho(\eta, \xi)$ есть косинус угла между η и ξ в соответствующем пространстве. Аналогично можно думать, что R^2 есть косинус между η и линейным пространством, натянутым на $\xi_1, \ldots, \xi_k$:

$$R^2 = \cos^2(\eta, \mathcal{L}(\xi_1, \ldots, \xi_k)).$$

Для коэффициента детерминации то же самое, только вместо случайных величин стоят вектора-признаки и косинус — это обычный косинус угла между векторами.

Возможна ситуация, когда $\cos^2(\eta, \mathcal{L}(\xi_1, \xi_2)) = 1 = R^2(\eta; \xi_1, \xi_2)$ — т.е. η лежит в $\mathcal{L}(\xi_1, \xi_2)$ (и предсказание абсолютно точно), но, тем не менее, $\mathrm{cor}(\xi_1, \eta) \approx 0$, $\mathrm{cor}(\xi_2, \eta) \approx 0$. Это возможно, если $\mathrm{cor}(\xi_1, \xi_2) \approx \pm 1$ (почти коллинеарны). ξ_1 называется «саппрессором» (suppressor) по отношению к ξ_2 (или наоборот). Подробнее, см https://stats.stackexchange.com/a/73876.

3.12.7. Как понять, что все хорошо

Если stepwise регрессия вперед и назад дает примерно одинаковые результаты и вторая производная одного знака (отрицат.), если нет супрессоров, нет плохих доверительных эллипсоидов, нет коэффициентов регрессии (перед стандартизованными признаками) больше 1.

Но самый хороший вариант, конечно, если регрессоры (почти) независимы. Если удается найти набор слабо зависимых признаков, которые предсказывают лишь немного хуже, чем полный набор, то это удача.

Также, если данные (регрессоры) должны собираться, то добавляются еще неформальные характеристики признаков — признаки должны легко и дешево собираться и не иметь много пропусков.

3.12.8. Заполнение пропусков

К стандартным вариантам casewise и pairwise добавляет вариант заполнения пропусков средним значением. Здесь такая опасность: если большое количество пропусков заполнить средними, то искусственно уменьшится дисперсия признаков и, тем самым, ширина доверительных интервалов для оценок параметров (увеличится значимость).

Есть еще варианты заполнения пропусков по регрессии на признаки с малым числом пропусков, но это нужно делать в ручном режиме.

3.13. Анализ аутлаеров

3.13.1. Matrix plot

Аутлаеров можно найти «на глаз» при помощи стандартного matrix plot данных.

3.13.2. Deleted residuals

можно применить технику кросс-валидации: удалить признак, построить модель, сравнить. Если индивид является аутлаером, то наблюдение y_i на нём «перетягивает» на себя регрессионную прямую. Тогда явно «большой» будет разница

$$r_i^{(i)} = \hat{y}_i^{(i)} - y_i$$

между $\hat{y}_i^{(i)}$ — значением регрессии на i-м индивиде без этого индивида и y_i — наблюдении на i-м индивиде. $r_i^{(i)}$ будет «большой» также по сравнению с $r_i = \hat{y}_i - y_i$. Напротив, если i-й индивид аутлаером не является, то будет справедливо приближенное равенство $r_i^{(i)} \approx r_i$, так что графиком $(r_i, r_i^{(i)})$ будет прямая. Deleted residuals всегда не меньше residuals, поэтому прямая $y = x$ не получится.

3.13.3. Studentized residuals

Справедливо

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \implies (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

откуда

$$\operatorname{cov}(\mathbf{y} - \hat{\mathbf{y}}) = \operatorname{cov}(\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})^{\mathsf{T}} \operatorname{cov} \mathbf{y} (\mathbf{I} - \mathbf{H}) = \sigma^2 (\mathbf{I} - \mathbf{H})^{\mathsf{T}} (\mathbf{I} - \mathbf{H}) = \sigma^2 (\mathbf{I} - \mathbf{H})$$

потому что $\mathbf{I} - \mathbf{H}$ — матрица проектора. Тогда,

$$\mathsf{D}(y_i - \hat{y}_i) = \mathsf{D}r_i = \sigma^2(1 - h_{ii}).$$

Как следствие, $\mathsf{D}\epsilon_i = \sigma^2 \geq \mathsf{D}r_i$.

Определение. h_{ii} — рычаг[5].

Чем больше i-й рычаг, тем меньше ошибка на i-м индивиде, так как он перетягивает на себя.

Определение. Стандартизированные остатки:

$$\frac{r_i}{\sqrt{\mathsf{D}r_i}} = \frac{r_i}{\sigma\sqrt{1 - h_{ii}}}.$$

Можно рассмотреть $\hat{\sigma}^{(i)}$ — оценку дисперсии без i-го индивида; тогда, при нормально распределенных ошибках наблюдения,

$$\frac{r_i}{\hat{\sigma}^{(i)}\sqrt{1 - h_{ii}}} \sim \mathrm{t}(n - m - 1)$$

(«-1» потому что меньше на одного индивида).

Замечание. Полученную величину можно сравнивать со «средним» значением рычага

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\operatorname{tr}\mathbf{H} = \frac{1}{n}\operatorname{rk}\mathbf{H} = \frac{k+1}{n}.$$

(как след идемпотентной матрицы, равный её рангу[6]: след есть сумма собственных чисел, однако у идемпотента два возможных собственных числа: 0 и 1, а кратность 1 в точности равна рангу).

[5] «Leverage».
[6] http://math.stackexchange.com/a/101515

3.13.4. Расстояние по Куку и расстояние Махаланобиса

Пусть $\hat{\mathbf{b}}^{(i)}$ — оценка без i-го индивида. Если расстояние между $\hat{\mathbf{b}}^{(i)}$ и $\hat{\mathbf{b}}$ «большое», то i-й индивид есть аутлаер:

$$r_M^2(\hat{\mathbf{b}}, \hat{\mathbf{b}}^{(i)}; \operatorname{cov}\hat{\mathbf{b}}) = (\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})^{\mathsf{T}} \operatorname{cov}^{-1}(\hat{\mathbf{b}})(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)}) = \frac{1}{\sigma^2}(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})\mathbf{X}^{\mathsf{T}}\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})$$

так что расстояние по Куку определяется как

$$\frac{(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})\mathbf{X}^{\mathsf{T}}\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})/m}{\hat{\sigma}^2}.$$

Расстояние по Куку показывает выбросы по отношению к регрессии (outliers всегда по отношению к чему-то, какой-то закономерности). Можно еще рассмотреть выбросы по отношению к распределению регрессоров (зависимая переменная тут не участвует). Это делается стандартным способом, через расстояние Махаланобиса в пространстве независимых признаков (регрессоров): если x_i — i-й индивид, $\bar{\mathbf{x}}$ — вектор средних, то аутлаером можно назвать индивида, для которого велико

$$r_M^2(x_i, \bar{\mathbf{x}}; \mathbf{S_{xx}}).$$

Правда, тут мы незаметно перешли к пониманию матрицы $\mathbf{X}$ как выборки из многомерного распределения. Если $\mathbf{X}$ — детерминированная матрица, то в этом смысле выбросов быть не может (так как нет закономерности).

Замечание. Если индивид аутлаер по Махаланобису, то $\mathbf{S_{xx}}$ оценивается неправильно (если понимать ее как выборочную ковариационную матрицу) и все значимости/доверительные интервалы становятся неправильными.

	Аутлаер по Куку	Не аутлаер по Куку
Аутлаер по Махаланобису	Далеко от линии регрессии	Далеко от $\bar{\mathbf{x}}$ на линии регрессии
Не аутлаер по Махаланобису	Близко к $\bar{\mathbf{x}}$ по одной координате и далеко по другой	Близко к $\bar{\mathbf{x}}$

3.14. Проверка правильности и выбор модели

- Если известно, что ошибки нормально распределены (например, в случае измерений прибора), то если остатки не имеют нормального распределения, то модель не является правильной.
- Если исходные данные имеют нелинейную зависимость, то и расположение остатков по линейной регрессии на графике будет отражать характер этой зависимости.
- Модель с наименьшим количеством параметров при прочих равных является предпочтительной, поэтому если заранее известно, что среднее 0, то свободный член из модели лучше удалить.

Замечание. В случае нормального распределения для всех случайных величин справедливо

$$\eta = \mathsf{E}(\eta \mid \xi_1, \ldots, \xi_i) + (\eta - \mathsf{E}(\eta \mid \xi_1, \ldots, \xi_i)).$$

Поэтому получается ортогональность остатков регрессии (регрессия линейна в силу нормальности). Значит все модели «верны» и можно среди них выбрать наилучшую.

3.15. Доверительные интервалы для предсказания

Пусть $(1, \mathbf{z})^\mathsf{T} \in \mathbb{R}^{k+1}$; тогда настоящее предсказание есть

$$\bar{\mathbf{y}} = \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} \mathbf{b},$$

а его оценка[7]

$$\hat{\mathbf{y}} = \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} \hat{\mathbf{b}}.$$

Эта оценка несмещенная, $\mathsf{E}\hat{\mathbf{y}} = \bar{\mathbf{y}}$. Можно показать, что её дисперсия есть

$$\mathsf{D}\hat{\mathbf{y}} = \sigma^2 \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n}(\mathbf{z} - \bar{\mathbf{x}})^\mathsf{T}\mathbf{S}_{\mathbf{xy}}^{-1}(\mathbf{z} - \bar{\mathbf{x}}),$$

частный случай чего выписывался в случае парной регрессии

$$\mathsf{D}\hat{\mathbf{y}} = \frac{\sigma^2}{n} + \frac{\sigma^2(x - \bar{\mathbf{x}})}{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})}.$$

Доверительным интервалом (оценки того, какое среднее значение будет на выходе, если на входе $\mathbf{z}$) будет

$$\bar{\mathbf{y}} \pm c_\gamma \,\mathrm{SE} = \bar{\mathbf{y}} \pm c_\gamma \sqrt{\mathsf{D}\hat{\mathbf{y}}} = \bar{\mathbf{y}} \pm c_\gamma \hat{\sigma} \sqrt{\begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}}, \quad c_\gamma \sim \mathrm{t}(n-m)$$

В данной модели $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$ — значение вообще, а $\mathbf{X}\mathbf{b}$ — среднее. Можно, поэтому, также построить ДИ для значения вообще:

$$\bar{\mathbf{y}} \pm c_\gamma \hat{\sigma} \sqrt{\underbrace{\mathsf{D}\epsilon}_{1} + \frac{1}{n} + \frac{1}{n}(\mathbf{z} - \bar{\mathbf{x}})^\mathsf{T}\mathbf{S}_{\mathbf{xy}}^{-1}(\mathbf{z} - \bar{\mathbf{x}})}, \quad c_\gamma \sim \mathrm{t}(n-m).$$

3.16. Сведение нелинейной модели к линейной

Существует три базовых модели, в которых функция регрессии линейная:

1. $\eta, \xi_1, \ldots, \xi_m$ нормальные.
2. $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$, $\mathsf{E}\boldsymbol{\epsilon} = \mathbf{0}$, $\operatorname{cov}\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$.
3. $y = a\xi + b + \epsilon$, ξ принимает всего два значения (возможно, как качественный признак).

Пусть

$$\eta = \phi(\xi_1, \ldots, \xi_k) + \epsilon$$

и ϕ — нелинейная функция.

- ϕ — многочлен. Можно свести к линейной, добавляя признаки $\xi, \xi^2, \ldots$ и для этих признаков строить модель.
- ξ — качественный признак. Можно ввести $k-1$ штук[8] фиктивных признаков со значениями $\{0, 1\}$ и для них строить модель. $A_1, \ldots, A_k$ — градации ϕ.

[7] Mean prediction.

[8] При добавлении вектора из единиц к k признакам получается вырожденная матрица.

3.17. Другие странные замечания

- $\eta = \phi(\xi) + \epsilon,\ \mathsf{E}\epsilon = 0, \epsilon \perp\!\!\!\perp \xi$. Если ϕ не линейная, то в ϵ войдет кусочек ξ и независимости не будет.
- Остатки всегда ортогональны т.к. проектор $\implies$ график — горизонтальная прямая всегда
- Графике $\hat{y}_i$ против $\hat{y}_i - y_i$ может быть наклонной прямой в случае pairwise MD deletion (и ковариационная матрица не соответствует данным).

4. Модификации линейной регрессии.

4.1. Взвешенная регрессия (Weighted Least Squares)

Пусть $\mathbf{W}$ — симметричная, положительно определенная матрица, тогда

$$\hat{\mathbf{b}}_W = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{y}$$

есть «взвешенная» оценка. При $\mathbf{W} = \mathbf{I}$, $\hat{\mathbf{b}}_W = \hat{\mathbf{b}}$, конечно.

Если $\mathsf{E}\boldsymbol{\epsilon} = \mathbf{0}$, то

$$\mathsf{E}\hat{\mathbf{b}}_W = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathsf{E}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}\mathbf{b} = \mathbf{b}$$

и оценка несмещенная.

- Если $\operatorname{cov}\boldsymbol{\epsilon} = \sigma^2\mathrm{I}$, то $\hat{\mathbf{b}}$ — BLUE и $\hat{\mathbf{b}}_W$ уже не лучшая.
- Если $\operatorname{cov}\boldsymbol{\epsilon} = \mathbf{C}$ (то есть шум не белый) то нужно подобрать $\mathbf{W}$ такую, что $\hat{\mathbf{b}}_W$ — BLUE. Это делается операцией отбеливания: пусть всё центрированное; тогда $\mathbf{C}^{-1/2}\boldsymbol{\epsilon}$ — центрированный и нормированный белый шум, и

  $$\underbrace{\mathbf{C}^{-1/2}\mathbf{y}}_{\tilde{\mathbf{y}}} = \underbrace{\mathbf{C}^{-1/2}\mathbf{X}}_{\tilde{\mathbf{X}}}\mathbf{b} + \mathbf{C}^{-1/2}\boldsymbol{\epsilon}$$

  откуда

  $$\operatorname{cov}(\mathbf{C}^{-1/2}\mathbf{y}) = \left(\mathbf{C}^{-1/2}\right)^\mathsf{T}\operatorname{cov}(\mathbf{X}\mathbf{b} + \boldsymbol{\epsilon})\mathbf{C}^{-1/2} = \mathbf{I}.$$

  Так как теперь шум белый, следующая оценка будет BLUE:

  $$\begin{aligned}\hat{\tilde{\mathbf{b}}} &= \tilde{\mathbf{X}}^{-}\tilde{\mathbf{y}} = ((\mathbf{C}^{-1/2}\mathbf{X})^\mathsf{T}(\mathbf{C}^{-1/2}\mathbf{X}))^{-1}\left(\mathbf{C}^{-1/2}\mathbf{X}\right)^\mathsf{T}\mathbf{C}^{-1/2}\mathbf{y} \\ &= (\mathbf{X}^\mathsf{T}\left(\mathbf{C}^{-1/2}\right)^\mathsf{T}\mathbf{C}^{-1/2}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\left(\mathbf{C}^{-1/2}\right)^\mathsf{T}\mathbf{C}^{-1/2}\mathbf{y}\end{aligned}$$

  Значит, следует положить $\mathbf{W} = \mathbf{C}^{-1}$.

  Для $\mathbf{W}$ итеративный процесс: берем начальное значение, находим коэффициент, оцениваем $\mathbf{C}$ и т.д.

Пример. Стандартный случай — измерения с разной точностью, откуда

$$\mathbf{C} = \begin{pmatrix}\sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_n^2\end{pmatrix}.$$

Наблюдениям, таким образом, придается разный вес — чем меньше точность наблюдения, тем больше σ_i^2 и меньший, соответственно, вес.

Замечание. $\mathbf{W}$ можно также назначить и руками.

4.2. Гребневая (Ridge) регрессия

Чтобы бороться с вырожденностью $\mathbf{R}_{\mathbf{xx}}$ в оценке $\hat{\boldsymbol{\beta}} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{R}_{\mathbf{xy}}$ рассматривают

$$\hat{\boldsymbol{\beta}} = (\mathbf{R}_{\mathbf{xx}} + \lambda\mathbf{I})^{-1}\mathbf{R}_{\mathbf{xy}}.$$

Получается смещенная оценка, но с меньшей дисперсией. Для поиска λ используют кросс-валидацию.

A. Свойства условного математического ожидания

1. $\mathsf{E}\{a\xi + b\theta \mid \eta\} = a\mathsf{E}\{\xi \mid \eta\} + b\mathsf{E}\{\theta \mid \eta\}$.
2. $\mathsf{E}\mathsf{E}\{\eta \mid \xi\} = \mathsf{E}\eta$.
3. $\xi \perp\!\!\!\perp \eta \implies \mathsf{E}\{\eta \mid \xi\} = \mathsf{E}\eta$.
4. $\eta = f(\xi) \implies \mathsf{E}\{\eta \mid \xi\} = \mathsf{E}\{f(\xi) \mid \xi\} = f(\xi)$.
5. $\mathsf{E}(\eta f(\xi) \mid f(\xi)) = f(\xi)\mathsf{E}\{\eta \mid \xi\}$.
6. $(\xi, \eta)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \mathsf{E}(\eta \mid \xi) = a\xi + b$.
   Замечание (Важное). Таким образом, если выборка нормальная, то зависимость линейная *всегда*.
7. $\operatorname{argmin}_{\hat{\eta} \in K = \{\phi(\xi)\}} \mathsf{E}(\eta - \hat{\eta})^2 = \mathsf{E}\{\eta \mid \xi\} = \hat{\eta}^*$.