УДК 519.22(075.8)
ББК 22.172я73
Б30

Печатается по решению
Редакционно-издательского совета
Белорусского государственного университета

Р е ц е н з е н т ы:
доктор физико-математических наук,
профессор *А. Д. Егоров*;
доктор физико-математических наук,
профессор *А. В. Лебедев*

Бахтин, В. И.
Б30 Введение в прикладную статистику : курс лекций. В 2 ч. Ч. 2. Методы прикладной статистики / В. И. Бахтин. – Минск : БГУ, 2012. – 99 с.
ISBN 978-985-518-633-6.

Во второй части курса лекций «Введение в прикладную статистику» содержатся разделы статистики, наиболее часто используемые в эконометрике: свойства многомерного нормального распределения, линейная регрессия и метод наименьших квадратов, главные компоненты, дискриминантный и кластерный анализ, дисперсионный анализ, временные ряды.
Предназначено для студентов математических специальностей.

УДК 519.22(075.8)
ББК 22.172я73

ISBN 978-985-518-633-6 (ч. 2)
ISBN 978-985-518-491-2

ПРЕДИСЛОВИЕ

В отличие от первой части курса лекций «Введение в прикладную статистику», где изложены основы классической математической статистики, во второй части рассматриваются методы прикладной статистики: линейная регрессия, метод главных компонент, дискриминантный и кластерный анализ, статистический анализ временных рядов, а также некоторые другие задачи прикладного характера. Все вышеперечисленные методы, так или иначе, базируются на свойствах многомерного нормального распределения. Поэтому изложение начинается с детального изучения последних.

Данную часть курса можно также рассматривать как математическое введение в *эконометрику* (тот раздел экономической науки, где занимаются статистическими методами обнаружения и исследования зависимостей между различными экономическими переменными). Отметим, что эконометрические задачи очень трудно (часто даже невозможно) решать «вручную», без применения специально разработанных для этой цели пакетов компьютерных программ. Поэтому в лекциях основное внимание уделяется математической (статистической) составляющей, а эмпирическая составляющая эконометрики (применение теоретических методов к конкретным экономическим задачам) должна изучаться во время лабораторного практикума, проводимого параллельно с чтением лекций.

Стоит отметить, что в прикладной статистике случайные векторы рассматриваются чаще, чем скалярные случайные величины. По этой причине во второй части всюду, где это возможно, используются геометрические подходы и терминология. Например, линейная регрессия всегда трактуется как ортогональная проекция на некоторое линейное подпространство, коэффициент корреляции (в том числе множественной и частной) — как косинус угла между двумя векторами евклидова (или гильбертова) пространства, матрица ковариаций — как метрический тензор (матрица Грама). При этом наиболее часто используемой теоремой в разделе о линейных регрессиях будет теорема Пифагора. Думаю, что такие геометрические определения легче для понимания, чем формальные.

Во второй части курса лекций содержатся главы с третьей по шестую и приложение. В третьей главе изучается многомерное нормальное распределение. Доказывается критерий независимости для набора нормально распределенных случайных величин, определяются условные нормальные распределения и метрика Махаланобиса, изучаются проекции и вращения изотропного нормального распределения, строятся и исследуются оценки разных параметров нормальных распределений, включая Z-статистику Фишера для коэффициента корреляции. В четвертой главе рассматривается линейная регрессия. В ней наряду с самой линейной регрессией определяются множественные и частные корреляции, для них строятся оценки, а затем описываются метод наименьших квадратов и модель множественной линейной регрессии. Пятая глава включает в себя несколько избранных тем из многомерного статистического анализа: метод главных компонент, кластерный и дискриминантный анализ, сравнение математических ожиданий при помощи T^2-статистики Хотеллинга, одномерный и многомерный дисперсионный анализ. В последней главе описываются различные типы временных рядов: стационарные, авторегрессии и скользящие средние, временные ряды с трендом. Анализируются методы идентификации типов временных рядов и оценки их параметров.

В приложении приведены используемые в курсе сведения из теории вероятностей и линейной алгебры: извлечение квадратного корня из положительной самосопряженной матрицы, матрицы Грама, многомерная центральная предельная теорема, линейные рекуррентные последовательности.

К сожалению, в односеместровом курсе лекций не представляется возможным отразить все темы, которые обычно включают в учебную дисциплину «Эконометрика». Например, в нем ничего не говорится о системах одновременных уравнений. С другой стороны, поскольку эти лекции предназначены студентам-математикам, в них число голословных утверждений сведено к минимуму, а большинство доказывается.

Как и в первой части, материал повышенной сложности набран в тексте мелким шрифтом. Символом «†» помечены заголовки тех параграфов, материал которых не используется в дальнейшем. Нумерация разделов сквозная в обеих частях.

Глава 3. МНОГОМЕРНОЕ НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

§ 22. Определение и свойства многомерного нормального распределения

Векторы x, t, μ, … мы всегда будем считать вектор-столбцами, а соответствующие вектор-строки будем обозначать x^*, t^*, μ^*, …

Определение. Случайный вектор $x \in \mathbb{R}^n$ *нормально распределен,* если он имеет характеристическую функцию

$$\varphi(t) = \mathsf{E}\big\{e^{it^*x}\big\} = e^{it^*\mu - \frac{1}{2}t^*\Sigma t}, \qquad t \in \mathbb{R}^n, \tag{22.1}$$

где $\mu \in \mathbb{R}^n$, а Σ — симметричная неотрицательно определенная матрица размерности $n \times n$. Соответствующее распределение вероятностей на пространстве $\mathbb{R}^n$ обозначается $\mathcal{N}_n(\mu, \Sigma)$. Нормальное распределение называют *невырожденным,* если $|\Sigma| \neq 0$. Очевидно, это условие равносильно тому, что матрица Σ положительно определена.

Подчеркнем, что из этого определения *не вытекает* существование нормального распределения. Оно будет доказано немного позже, в предложении 22.4. В предположении, что нормальное распределение существует, его единственность (при заданных параметрах μ и Σ) следует из того, что разным распределениям вероятностей отвечают разные характеристические функции.

Введем обозначения для компонент векторов x, t, μ и матрицы Σ:

$$x = (x_1, \ldots, x_n), \qquad t = (t_1, \ldots, t_n),$$
$$\mu = (\mu_1, \ldots, \mu_n), \qquad \Sigma = (\sigma_{kl})_{k,l=1}^n.$$

Предложение 22.1. *Если случайный вектор $x \in \mathbb{R}^n$ имеет нормальное распределение $\mathcal{N}_n(\mu, \Sigma)$, то тогда $\mu = \mathsf{E}x$, $\Sigma = \mathrm{Cov}\{x, x\}$, а центральные моменты четвертого порядка имеют вид*

$$\mathsf{E}\big\{(x_k - \mu_k)(x_l - \mu_l)(x_p - \mu_p)(x_q - \mu_q)\big\} = \\ = \sigma_{kl}\sigma_{pq} + \sigma_{kp}\sigma_{lq} + \sigma_{kq}\sigma_{lp}. \tag{22.2}$$

Доказательство. Умножим равенства (22.1) на $e^{-it^*\mu}$:

$$\mathsf{E}\big\{e^{it^*(x-\mu)}\big\} = e^{-\frac{1}{2}t^*\Sigma t}. \tag{22.3}$$

Очевидно,

$$t^*(x-\mu) = \sum_{r=1}^{n} t_r(x_r - \mu_r), \qquad t^*\Sigma t = \sum_{s,r=1}^{n} t_s \sigma_{sr} t_r.$$

Продифференцируем равенство (22.3) по переменной t_k:

$$\mathsf{E}\big\{e^{it^*(x-\mu)} i(x_k - \mu_k)\big\} = -\frac{1}{2}\sum_{r=1}^{n}\big(\sigma_{kr}t_r + t_r\sigma_{rk}\big)e^{-\frac{1}{2}t^*\Sigma t} =$$

$$= -\sum_{r=1}^{n}\sigma_{kr}t_r e^{-\frac{1}{2}t^*\Sigma t}. \tag{22.4}$$

При $t = 0$ получаем, что $\mathsf{E}\{x_k - \mu_k\} = 0$ и, следовательно, $\mathsf{E}x = \mu$. Далее, продифференцируем равенство (22.4) по переменной t_l:

$$\mathsf{E}\big\{-e^{it^*(x-\mu)}(x_k - \mu_k)(x_l - \mu_l)\big\} =$$

$$= -\sigma_{kl}e^{-\frac{1}{2}t^*\Sigma t} + \left(\sum_{r=1}^{n}\sigma_{kr}t_r\right)\left(\sum_{r=1}^{n}\sigma_{lr}t_r\right)e^{-\frac{1}{2}t^*\Sigma t}. \tag{22.5}$$

При $t = 0$ из (22.5) получаем

$$\mathrm{Cov}\{x_k, x_l\} = \mathsf{E}\big\{(x_k - \mu_k)(x_l - \mu_l)\big\} = \sigma_{kl}.$$

Значит, $\mathrm{Cov}\{x, x\} = \Sigma$. Наконец, продифференцируем формулу (22.5) еще по t_p и по t_q при $t = 0$. Тогда у нас получится (22.2). □

Предложение 22.2. *Пусть случайный вектор $x \in \mathbb{R}^n$ имеет нормальное распределение $\mathcal{N}_n(\mu, \Sigma)$. Тогда случайный вектор $y = Cx + b$, где $C\colon \mathbb{R}^n \to \mathbb{R}^m$ — любое линейное отображение, имеет нормальное распределение $\mathcal{N}_m(C\mu + b, C\Sigma C^*)$.*

Доказательство. Вычислим характеристическую функцию y:

$$\varphi_y(\tau) = \mathsf{E}\big\{e^{i\tau^*y}\big\} = \mathsf{E}\big\{e^{i\tau^*(Cx+b)}\big\} = \mathsf{E}\big\{e^{i(C^*\tau)^*x}\big\}e^{i\tau^*b} = \varphi_x(C^*\tau)e^{i\tau^*b} =$$

$$= e^{i(C^*\tau)^*\mu - \frac{1}{2}(C^*\tau)^*\Sigma C^*\tau}e^{i\tau^*b} = e^{i\tau^*(C\mu+b) - \frac{1}{2}\tau^*C\Sigma C^*\tau}.$$

Последнее выражение совпадает с характеристической функцией распределения $\mathcal{N}_m(C\mu + b, C\Sigma C^*)$. $\square$

Следствие 22.3. *Если вектор* $x' = (x_{i_1}, \ldots, x_{i_m})$ *составлен из нескольких координат случайного вектора* $x \in \mathbb{R}^n$ *с нормальным распределением* $\mathcal{N}_n(\mu, \Sigma)$, *то он имеет нормальное распределение* $\mathcal{N}_m(\mu', \Sigma')$, *где* $\mu' = \mathsf{E}x' = (\mu_{i_1}, \ldots, \mu_{i_m})$ *и* $\Sigma' = \operatorname{Cov}\{x', x'\} = (\sigma_{i_k i_l})_{k,l=1}^m$.

Доказательство. Случайный вектор x' линейно зависит от x и по предложению 22.2 нормально распределен. Математические ожидания и ковариации его компонент совпадают с математическими ожиданиями и ковариациями тех же самых компонент вектора x. $\square$

Задача 1. Докажите, что для *любого* случайного вектора x если $y = Cx + b$, то $\mathsf{E}y = C\mathsf{E}x + b$ и $\operatorname{Cov}\{y, y\} = C\operatorname{Cov}\{x, x\}C^*$.

Задача 2. Приведите пример случайного вектора, у которого все координаты нормально распределены, но сам он не нормально распределен.

Предложение 22.4. *Если* $|\Sigma| \neq 0$, *то нормальное распределение* $\mathcal{N}_n(\mu, \Sigma)$ *существует и имеет плотность*

$$p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^*\Sigma^{-1}(x-\mu)\right\}, \qquad x \in \mathbb{R}^n.$$

Доказательство. Пусть у случайного вектора $y = (y_1, \ldots, y_n)$ все компоненты имеют распределение $\mathcal{N}_1(0, 1)$ и независимы. Тогда его плотность вероятности равна

$$f(y) = \prod_{j=1}^n f(y_j) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi}} e^{-y_j^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}y^*y},$$

а характеристическая функция имеет вид

$$\varphi_y(\tau) = \mathsf{E}\left\{e^{i\tau^*y}\right\} = \mathsf{E}\left\{\prod_{j=1}^n e^{i\tau_j y_j}\right\} = \prod_{j=1}^n \mathsf{E}\left\{e^{i\tau_j y_j}\right\} = \prod_{j=1}^n e^{-\tau_j^2/2} = e^{-\frac{1}{2}\tau^*\tau}.$$

Отсюда вытекает, что случайный вектор y имеет нормальное распределение $\mathcal{N}_n(0, I)$, где I — единичная матрица размерности $n \times n$.

Пусть $\Sigma^{1/2}$ — симметричная положительно определенная матрица, для которой справедливо равенство $(\Sigma^{1/2})^2 = \Sigma$ (она строится в приложении, п. III). Сделаем линейную замену переменных $x = \mu + \Sigma^{1/2}y$.

Тогда $y = \Sigma^{-1/2}(x - \mu)$. По предложению 22.2 случайная величина x нормально распределена с математическим ожиданием μ и матрицей ковариаций $\Sigma^{1/2} I \Sigma^{1/2} = \Sigma$. По формуле замены переменных в интеграле плотность вероятности для x равна

$$p(x) = f(y) \left| \det\left(\frac{dy}{dx} \right) \right| = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-\mu)^* \Sigma^{-1}(x-\mu)} \left| \Sigma^{-1/2} \right|. \quad \square$$

Задача 3. Как устроено нормальное распределение с вырожденной матрицей ковариаций Σ?

Подсказка: выберите систему координат, в которой она становится диагональной.

Теорема 22.5 (критерий независимости). *Допустим, что вектор $x = (x_1, x_2, \ldots, x_k)$ составлен из компонент $x_i \in \mathbb{R}^{n_i}$, $i = 1, \ldots, k$, и имеет нормальное распределение. Тогда эти компоненты независимы в том и только том случае, когда они некоррелированы.*

Доказательство. Независимые случайные величины некоррелированы. Наоборот, если $\Sigma_{ij} = \mathrm{Cov}\{x_i, x_j\} = 0$ при $i \neq j$, то матрица $\Sigma = \mathrm{Cov}\{x, x\}$ имеет блочно-диагональный вид

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \Sigma_{kk} \end{pmatrix},$$

и характеристическая функция случайного вектора x представляется как произведение характеристических функций векторов $x_1, \ldots, x_k$:

$$\varphi_x(t) = e^{it^*\mu - \frac{1}{2}t^*\Sigma t} = \prod_{j=1}^{k} e^{it_j^*\mu_j - \frac{1}{2}t_j^*\Sigma_{jj}t_j},$$

где $\mu = \mathsf{E}x$ и $\mu_j = \mathsf{E}x_j$. Отсюда следует независимость $x_1, \ldots, x_k$. $\square$

§ 23. Метрика Махаланобиса и распределение χ^2

Для невырожденного нормального распределения $\mathcal{N}_n(\mu, \Sigma)$ вводят *метрику Махаланобиса*

$$\rho(x, y) = \sqrt{(x-y)^*\Sigma^{-1}(x-y)}, \qquad x, y \in \mathbb{R}^n.$$

Она порождается скалярным произведением $(x, y) = x^*\Sigma^{-1}y$.

Предложение 23.1. *Если случайный вектор x имеет нормальное распределение $\mathcal{N}_n(\mu, \Sigma)$, где $|\Sigma| \neq 0$, то случайная величина $\rho^2(x, \mu) = (x - \mu)^*\Sigma^{-1}(x - \mu)$ распределена как χ^2_n.*

Доказательство. Сделаем линейную замену $z = \Sigma^{-1/2}(x - \mu)$. По предложению 22.2 случайный вектор z имеет нормальное распределение $\mathcal{N}_n(0, I)$, в котором $I = \Sigma^{-1/2}\Sigma\Sigma^{-1/2}$ — единичная матрица. По предложению 22.5 координаты случайного вектора $z = (z_1, \ldots, z_n)$ независимы и имеют распределение $\mathcal{N}_1(0, 1)$. Поэтому

$$\rho^2(x, \mu) = (x - \mu)^*\Sigma^{-1/2}\Sigma^{-1/2}(x - \mu) = z^*z = \sum_{i=1}^{n} z_i^2 = \chi^2_n. \quad \square$$

Эллипсоидом рассеяния радиуса R принято называть множество $V_R = \{x \mid \rho(x, \mu) \leqslant R\}$. Из предыдущего предложения следует, что вероятность попадания случайного вектора x в эллипсоид рассеяния V_R равняется $F_{\chi^2_n}(R^2)$, где $F_{\chi^2_n}$ — функция распределения случайной величины χ^2_n.

Теперь мы можем доказать теорему 18.1 об асимптотике распределения статистики χ^2 из критерия согласия Пирсона. Напомним ее формулировку.

Пусть имеется случайная величина x с функцией распределения $F_0(x)$. Вещественная прямая разбивается на m частей точками $-\infty = b_0 < b_1 < \ldots < b_{m-1} < b_m = +\infty$. По этим точкам строятся полуинтервалы $\Delta_k = (b_{k-1}, b_k]$. Вероятность попадания x в полуинтервал Δ_k равна $p_k = F_0(b_k) - F_0(b_{k-1})$. Предполагается, что все $p_k > 0$.

Рассмотрим выборку $X = (x_1, \ldots, x_N)$ из распределения $F_0(x)$. Пусть ν_k — число выборочных значений, попавших в Δ_k. Статистика χ^2 Пирсона имеет вид

$$\chi^2(X) = \sum_{k=1}^{m} \frac{(\nu_k - Np_k)^2}{Np_k} = \sum_{k=1}^{m} y_k^2, \qquad \text{где} \quad y_k = \frac{\nu_k - Np_k}{\sqrt{Np_k}}.$$

Требуется доказать, что при $N \to \infty$ ее распределение сходится к распределению случайной величины χ^2_{m-1}.

Определим вектор-функцию

$$\xi(x) = \big(\xi_1(x), \ldots, \xi_m(x)\big), \qquad \text{где} \quad \xi_k(x) = \begin{cases} 1, & \text{если } x \in \Delta_k, \\ 0, & \text{если } x \notin \Delta_k. \end{cases} \tag{23.1}$$

По построению случайные векторы $\xi(x_1)$, ..., $\xi(x_N)$ независимы и одинаково распределены. Найдем параметры их распределения:

$$\mathsf{E}\{\xi_k(x)\} = P\{\xi_k(x) = 1\} \cdot 1 + P\{\xi_k(x) = 0\} \cdot 0 = p_k,$$

$$\mathrm{Cov}\{\xi_k(x), \xi_l(x)\} = \mathsf{E}\{\xi_k(x)\xi_l(x)\} - \mathsf{E}\{\xi_k(x)\}\mathsf{E}\{\xi_l(x)\} = p_k\delta_{kl} - p_kp_l$$

(мы воспользовались тем, что $\xi_k^2(x) = \xi_k(x)$ и $\xi_k(x)\xi_l(x) = 0$ при $k \neq l$).

Из (23.1) следует, что $\xi_k(x_1) + \ldots + \xi_k(x_N) = \nu_k$. Поэтому

$$y_k = \sum_{t=1}^{N} \frac{\xi_k(x_t) - p_k}{\sqrt{Np_k}}, \qquad \mathsf{E}\, y_k = 0,$$

$$\mathrm{Cov}\{y_k, y_l\} = \sum_{t=1}^{N} \mathrm{Cov}\left\{\frac{\xi_k(x_t)}{\sqrt{Np_k}}, \frac{\xi_l(x_t)}{\sqrt{Np_l}}\right\} = \delta_{kl} - \sqrt{p_k}\sqrt{p_l}.$$

Рассмотрим линейное подпространство $L \subset \mathbb{R}^m$ вида

$$L = \bigl\{y = (y_1, \ldots, y_m) \bigm| \sqrt{p_1}\, y_1 + \ldots + \sqrt{p_m}\, y_m = 0\bigr\}.$$

Оно содержит случайный вектор y с координатами $y_k = (\nu_k - Np_k)\big/\sqrt{Np_k}$. Возьмем любые два вектора $\alpha, \beta \in \mathbb{R}^m$. Для них

$$\mathrm{Cov}\{\alpha^* y, \beta^* y\} = \sum_{k=1}^{m}\sum_{l=1}^{m} \alpha_k \beta_l \,\mathrm{Cov}\{y_k, y_l\} = \sum_{k=1}^{m} \alpha_k \beta_k - \sum_{k=1}^{m} \alpha_k \sqrt{p_k} \sum_{l=1}^{m} \beta_l \sqrt{p_l}.$$

В частности, если $\alpha, \beta \in L$, то тогда

$$\mathrm{Cov}\{\alpha^* y, \beta^* y\} = \sum_{k=1}^{m} \alpha_k \beta_k = \alpha^* \beta. \tag{23.2}$$

Выберем в подпространстве L любой ортонормированный базис $e_1, \ldots, e_{m-1}$. По центральной предельной теореме (см. приложение, п. I) распределение случайного вектора y сходится к нормальному. Значит, распределение случайного вектора $(e_1^* y, \ldots, e_{m-1}^* y)$ тоже сходится к нормальному. Это предельное распределение в силу (23.2) совпадает с $\mathcal{N}_{m-1}(0, I)$. Отсюда вытекает, что распределение случайной величины

$$\chi^2(X) = \sum_{k=1}^{m} y_k^2 = \sum_{k=1}^{m-1} (e_k^* y)^2$$

сходится к распределению χ^2_{m-1}. $\square$

§ 24. Условное нормальное распределение

Пусть составная случайная величина $z = (x, y)$ имеет нормальное распределение $\mathcal{N}_{n+m}(\mu, \Sigma)$. Мы здесь считаем, что $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ и, соответственно,

$$\mu = \mathsf{E}\, z = (\mu_x, \mu_y), \qquad \text{где } \ \mu_x = \mathsf{E}\, x, \ \ \mu_y = \mathsf{E}\, y;$$

$$\Sigma = \mathrm{Cov}\{z, z\} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \qquad \text{где } \ \Sigma_{xy} = \Sigma_{yx}^* = \mathrm{Cov}\{x, y\}.$$

Теорема 24.1. *Если распределение* $\mathcal{N}_{n+m}(\mu, \Sigma)$ *невырождено, то условное распределение случайного вектора* y *при фиксированном* x *есть нормальное распределение* $\mathcal{N}_m(\mu_{y|x}, \Sigma_{yy|x})$, *в котором*

$$\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x), \tag{24.1}$$

$$\Sigma_{yy|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{24.2}$$

Отсюда видно, что матрица ковариаций условного распределения случайной величины y при фиксированном x *не зависит* от x.

Доказательство. Рассмотрим случайный вектор

$$y' = y - \mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x).$$

Пара (x, y') линейно зависит от z и потому нормально распределена. При этом

$$\mathsf{E}y' = 0, \tag{24.3}$$

$$\operatorname{Cov}\{y', x\} = \mathsf{E}\{y'(x - \mu_x)^*\} = \Sigma_{yx} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xx} = 0, \tag{24.4}$$

$$\begin{aligned}\operatorname{Cov}\{y', y'\} = \mathsf{E}\big\{\big((y - \mu_y) - \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)\big) \times \\ \times \big((y - \mu_y)^* - (x - \mu_x)^*\Sigma_{xx}^{-1}\Sigma_{xy}\big)\big\} = \\ = \Sigma_{yy} - 2\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} + \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xx}\Sigma_{xx}^{-1}\Sigma_{xy} = \Sigma_{yy|x}.\end{aligned} \tag{24.5}$$

Из (24.4) следует, что случайная величина y' не зависит от x. Поэтому ее условное распределение совпадает с безусловным, а последнее в силу (24.3) и (24.5) имеет вид $\mathcal{N}_m(0, \Sigma_{yy|x})$. При фиксированном x величина $\mu_{y|x}$ является константой, и условное распределение y получается из распределения y' сдвигом на эту константу. □

Следствие 24.2. *В условиях предыдущей теоремы* $\mu_{y|x} = \mathsf{E}\{y\,|\,x\}$.

Следствие 24.3. *Величина* $\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)$ *является наилучшим (в среднеквадратичном) приближением* y *по* x: *для любого борелевского отображения* $f\colon \mathbb{R}^n \to \mathbb{R}^m$ *выполняется неравенство*

$$\mathsf{E}\big\{|y - \mu_{y|x}|^2\big\} \leqslant \mathsf{E}\big\{|y - f(x)|^2\big\}.$$

Доказательство. Это одно из свойств условного математического ожидания. □

§ 25. Проекции и вращения нормальных распределений

Лемма 25.1 (о проекциях). *Пусть случайный вектор $X \in \mathbb{R}^N$ распределен по нормальному закону $\mathcal{N}_N(\mu, \sigma^2 I)$. Пусть в $\mathbb{R}^N$ выбраны два линейных подпространства L_1, L_2, которые ортогональны друг другу и имеют нулевое пересечение, а P_1 и P_2 — ортогональные проекторы на L_1 и на L_2 соответственно. Тогда случайные векторы P_1X и P_2X взаимно независимы. Если при этом выполняется равенство $P_1\mu = 0$, то случайная величина $|P_1X|^2$ распределена как $\sigma^2\chi^2_n$, где $n = \dim L_1$. А если к тому же $P_2\mu = 0$, то случайная величина*

$$F = \frac{m}{n}\frac{|P_1X|^2}{|P_2X|^2},$$

где $m = \dim L_2$, имеет распределение Фишера $F_{n,m}$.

Доказательство. Выберем в подпространстве L_1 ортонормированный базис $e_1, \ldots, e_n$, а в подпространстве L_2 ортонормированный базис $e_{n+1}, \ldots, e_{n+m}$. Определим случайные величины $y_i = e_i^* X$. По предложению 22.2 они нормально распределены и имеют ковариации

$$\mathrm{Cov}\{y_i, y_j\} = \mathsf{E}\{e_i^*(X-\mu)(X-\mu)^* e_j\} = e_i^*(\sigma^2 I)e_j = \delta_{ij}\sigma^2.$$

Значит, все они независимы в совокупности, а их дисперсии равны σ^2. Очевидно,

$$P_1X = \sum_{i=1}^{n} y_i e_i, \qquad P_2X = \sum_{i=n+1}^{n+m} y_i e_i,$$

$$|P_1X|^2 = y_1^2 + \ldots + y_n^2, \qquad |P_2X|^2 = y_{n+1}^2 + \ldots + y_{n+m}^2.$$

Отсюда следует, что случайные векторы P_1X и P_2X независимы. Если $P_1\mu = 0$, то случайные величины $y_1, \ldots, y_n$ имеют одинаковое распределение $\mathcal{N}_1(0, \sigma^2)$. Поэтому величина $|P_1X|^2$ распределена как $\sigma^2\chi^2_n$. Если к тому же $P_2\mu = 0$, то совершенно аналогично случайная величина $|P_2X|^2$ распределена как $\sigma^2\chi^2_m$, а отношение

$$\frac{m}{n}\frac{|P_1X|^2}{|P_2X|^2} = \frac{m}{n}\frac{\chi^2_n}{\chi^2_m}$$

по определению имеет распределение Фишера $F_{n,m}$. $\square$

Лемма 25.2 (о вращении). *Пусть случайные векторы $x_1, \ldots, x_N$ из $\mathbb{R}^n$ независимы и имеют нормальные распределения с одинаковой матрицей ковариаций Σ (при этом их математические ожидания могут быть какими угодно), и пусть матрица $C = (c_{tk})_{t,k=1}^N$ ортогональна. Тогда случайные векторы $y_t = \sum_{k=1}^N c_{tk}x_k$, $t = 1, \ldots, N$, тоже независимы и имеют нормальные распределения с матрицей ковариаций Σ.*

Доказательство. Случайные векторы $x_1, \ldots, x_N$ нормально распределены в совокупности. Набор векторов $y_1, \ldots, y_N$ получается из $x_1, \ldots, x_N$ путем линейного преобразования и поэтому тоже имеет совокупное нормальное распределение. В частности, все y_t нормально распределены. Вычислим ковариации y_t и y_s:

$$\operatorname{Cov}\{y_t, y_s\} = \operatorname{Cov}\left\{\sum_{k=1}^N c_{tk}x_k, \sum_{l=1}^N c_{sl}x_l\right\} = \sum_{k=1}^N\sum_{l=1}^N c_{tk}c_{sl}\operatorname{Cov}\{x_k, x_l\} =$$

$$= \sum_{k=1}^N\sum_{l=1}^N c_{tk}c_{sl}\delta_{kl}\Sigma = \sum_{k=1}^N c_{tk}c_{sk}\Sigma = \delta_{ts}\Sigma.$$

Отсюда следует, что векторы $y_1, \ldots, y_N$ независимы и каждый из них имеет матрицу ковариаций Σ. $\square$

Определение. Если случайные векторы $z_1, \ldots, z_N \in \mathbb{R}^n$ независимы и имеют нормальное распределение $\mathcal{N}_n(0, \Sigma)$, то вероятностное распределение случайной матрицы $W = \sum_{t=1}^N z_t z_t^*$ называется *распределением Уишарта* с N степенями свободы и обозначается $\mathcal{W}_n(\Sigma, N)$.

Матрица W имеет размерность $n \times n$. Если $z_t = (z_{t1}, \ldots, z_{tn})$, то она состоит из элементов $w_{ij} = \sum_{t=1}^N z_{ti}z_{tj}$. При $n = 1$ матрица ковариаций Σ состоит из одного элемента $\sigma^2 = \mathsf{D}\, z_t$, и матрица W тоже содержит единственный элемент $w = \sum_{t=1}^N z_t^2$. Значит, распределение $\mathcal{W}_1(\Sigma, N)$ совпадает с распределением случайной величины $\sigma^2\chi_N^2$.

Определение. Если случайные матрицы A и B независимы и имеют распределения Уишарта $\mathcal{W}_n(\Sigma, N)$ и $\mathcal{W}_n(\Sigma, M)$, причем $|\Sigma| \neq 0$ и $N \geqslant n$, то случайная величина

$$\lambda = \lambda_n(N, M) = \frac{|A|}{|A + B|}$$

называется *статистикой Уилкса*.

Вопрос. Зачем здесь нужно условие $N \geqslant n$?

Предложение 25.3. *Вероятностное распределение статистики Уилкса не зависит от матрицы ковариаций* Σ.

Доказательство. По определению распределения Уишарта

$$A = \sum_{t=1}^{N} z_t z_t^*, \qquad B = \sum_{t=N+1}^{N+M} z_t z_t^*,$$

где $z_1, \ldots, z_{N+M}$ — независимые случайные векторы с распределением $\mathcal{N}_n(0, \Sigma)$. Рассмотрим случайные векторы $y_t = \Sigma^{-1/2} z_t$. Они независимы и имеют распределение $\mathcal{N}_n\big(0, \Sigma^{-1/2}\Sigma\Sigma^{-1/2}\big) = \mathcal{N}_n(0, I)$. Выразим через них статистику Уилкса:

$$\lambda = \frac{|A|}{|A+B|} = \frac{\left|\Sigma^{1/2}\big(\sum_{t=1}^{N} y_t y_t^*\big)\Sigma^{1/2}\right|}{\left|\Sigma^{1/2}\big(\sum_{t=1}^{N+M} y_t y_t^*\big)\Sigma^{1/2}\right|} = \frac{\left|\sum_{t=1}^{N} y_t y_t^*\right|}{\left|\sum_{t=1}^{N+M} y_t y_t^*\right|}.$$

Отсюда видно, что распределение λ не зависит от Σ. $\square$

Лемма 25.4 (Андерсон, [2]). *Пусть*

1) задана случайная матрица $X = (x_{tj})$ *размерности* $N \times n$;

2) все ее строки $x_t = (x_{t1}, \ldots, x_{tn})$ *независимы и имеют нормальные распределения с одной и той же матрицей ковариаций* Σ;

3) в $\mathbb{R}^N$ *заданы два ортогональных проектора* $P_1 : \mathbb{R}^N \to L_1$ *и* $P_2 : \mathbb{R}^N \to L_2$ *на линейные подпространства* $L_1, L_2 \subset \mathbb{R}^N$, *которые ортогональны друг другу и имеют нулевое пересечение.*

Тогда

а) случайные матрицы $P_1 X$ *и* $P_2 X$ *взаимно независимы*;

б) если $\mathsf{E}\{P_1 X\} = 0$, *то матрица* $X^* P_1 X$ *имеет распределение* $\mathcal{W}_n(\Sigma, k)$, *где* $k = \dim L_1$;

в) если $\mathsf{E}\{P_1 X\} = \mathsf{E}\{P_2 X\} = 0$ *и* $|\Sigma| \neq 0$, *то отношение*

$$\lambda = \frac{|X^* P_1 X|}{|X^*(P_1 + P_2)X|} \tag{25.1}$$

является статистикой Уилкса $\lambda_n(k, l)$, *где* $k = \dim L_1$ *и* $l = \dim L_2$.

Доказательство. Предположим вначале, что L_1 и L_2 — это координатные подпространства:

$$L_1 = \big\{(y_1, \ldots, y_k, 0, \ldots, 0) \in \mathbb{R}^N\big\},$$

$$L_2 = \big\{(0, \ldots, 0, y_{k+1}, \ldots, y_{k+l}, 0, \ldots, 0) \in \mathbb{R}^N\big\}.$$

Тогда у матрицы $P_1 X$ первые k строк совпадают с $x_1, \ldots, x_k$, а остальные строки нулевые, в то время как у матрицы $P_2 X$ строки с номерами $k+1, \ldots, k+l$ совпадают с $x_{k+1}, \ldots, x_{k+l}$, а остальные нулевые. Значит, матрицы $P_1 X$ и $P_2 X$ состоят из независимых наборов строк и потому независимы.

По построению матрица $W = X^* P_1 X$ состоит из элементов $w_{ij} = \sum_{t=1}^{k} x_{ti} x_{tj}$. Значит, она совпадает с матрицей $\sum_{t=1}^{k} x_t^* x_t$. Если при этом $\mathsf{E}\{P_1 X\} = 0$, то все строки $x_1, \ldots, x_k$ имеют нормальное распределение $\mathcal{N}_n(0, \Sigma)$, а матрица W имеет распределение Уишарта $\mathcal{W}_n(\Sigma, k)$. Если еще $\mathsf{E}\{P_2 X\} = 0$, то совершенно аналогично матрица $X^* P_2 X$ имеет распределение $\mathcal{W}_n(\Sigma, l)$, а отношение (25.1) является статистикой Уилкса $\lambda_n(k, l)$.

В общем случае определим в пространстве $\mathbb{R}^N$ ортогональный поворот C, который переводит L_1 и L_2 в координатные подпространства. По лемме 25.2 строки матрицы $Y = CX$ независимы и имеют нормальные распределения с одной и той же матрицей ковариаций Σ. В этих условиях для матрицы Y и подпространств $CL_1, CL_2 \subset \mathbb{R}^N$ лемма уже доказана.

Из линейной алгебры известно, что ортогональный проектор на подпространство CL_i имеет вид CP_iC^* (где P_i — проектор на L_i). В силу уже доказанного утверждения а) для координатных подпространств матрицы $(CP_1C^*)Y$ и $(CP_2C^*)Y$ независимы. Домножая их слева на C^*, получаем матрицы P_1X и P_2X. Значит, они тоже независимы. Если $\mathsf{E}\{P_1X\} = 0$, то $\mathsf{E}\{(CP_1C^*)Y\} = 0$, и в силу утверждения б) для координатных подпространств матрица $Y^*(CP_1C^*)Y = X^*P_1X$ имеет распределение $\mathcal{W}_n(\Sigma, k)$. Аналогично, если $\mathsf{E}\{P_2X\} = 0$, то матрица X^*P_2X имеет распределение $\mathcal{W}_n(\Sigma, l)$, а отношение (25.1) является статистикой Уилкса. □

§ 26. Выборочные среднее и ковариации многомерного нормального распределения

Рассмотрим выборку $X = (x_1, \ldots, x_N)$ из нормального распределения $\mathcal{N}_n(\mu, \Sigma)$. Вычислим для нее выборочное среднее и выборочную матрицу ковариаций

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t, \qquad \hat{\Sigma} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(x_t - \bar{x})^*.$$

Теорема 26.1. *Если $X = (x_1, \ldots, x_N)$ — выборка из нормального распределения, то оценки $\bar{x}$ и $\hat{\Sigma}$ независимы. При этом величина $\bar{x}$ распределена по нормальному закону $\mathcal{N}_n\big(\mu, \frac{1}{N}\Sigma\big)$, а матрица $N\hat{\Sigma}$ имеет распределение Уишарта $\mathcal{W}_n(\Sigma, N-1)$.*

Следствие 26.2. *В одномерном случае для нормального распределения $\mathcal{N}_1(\mu, \sigma^2)$ выборочное среднее $\bar{x}$ и выборочная дисперсия $\hat{\sigma}^2$ независимы, а статистика $N\hat{\sigma}^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2$ распределена как $\sigma^2\chi^2_{N-1}$.*

Доказательство. Построим ортогональную матрицу $C = (c_{tk})$ размерности $N \times N$, в которой все элементы первой строки имеют вид $c_{1k} = 1/\sqrt{N}$, а элементы других строк выбираются произвольно, лишь бы матрица была ортогональной. Из того, что первая строка ортогональна остальным, вытекают равенства

$$\sum_{k=1}^{N} c_{tk} = \sqrt{N} \sum_{k=1}^{N} c_{1k} c_{tk} = 0, \qquad t \geqslant 2.$$

Рассмотрим случайные векторы

$$z_t = \sum_{k=1}^{N} c_{tk} x_k, \qquad t = 1, \ldots, N.$$

В силу леммы 25.2 они независимы, нормально распределены, имеют одну и ту же матрицу ковариаций Σ и математические ожидания

$$\mathsf{E}\, z_t = \sum_{k=1}^{N} c_{tk} \mu = \begin{cases} \sqrt{N}\mu, & t = 1, \\ 0, & t > 1. \end{cases}$$

Очевидно, $z_1 = \sqrt{N}\,\bar{x}$. Поэтому $\bar{x}$ имеет распределение $\mathcal{N}_n\big(\mu, \frac{1}{N}\Sigma\big)$. Далее,

$$\sum_{t=1}^{N} z_t z_t^* = \sum_{t=1}^{N} \bigg(\sum_{k=1}^{N} c_{tk} x_k\bigg) \bigg(\sum_{l=1}^{N} c_{tl} x_l\bigg)^* =$$

$$= \sum_{k,l=1}^{N} \sum_{t=1}^{N} c_{tk} c_{tl} x_k x_l^* = \sum_{k,l=1}^{N} \delta_{kl} x_k x_l^* = \sum_{t=1}^{N} x_t x_t^*.$$

Следовательно,

$$N\hat{\Sigma} = \sum_{t=1}^{N} (x_t - \bar{x})(x_t - \bar{x})^* = \sum_{t=1}^{N} x_t x_t^* - N\bar{x}\bar{x}^* = \sum_{t=2}^{N} z_t z_t^*. \tag{26.1}$$

Отсюда видно, что матрица $N\hat{\Sigma}$ имеет распределение $\mathcal{W}_n(\Sigma, N-1)$ и не зависит от $\bar{x}$. $\square$

Оценка $\bar{x}$ для μ несмещенная. Из формулы (26.1) видно, что оценка $\hat{\Sigma}$ для Σ является смещенной, а несмещенной оценкой будет

$$S = \frac{N\hat{\Sigma}}{N-1} = \frac{1}{N-1} \sum_{t=1}^{N} (x_t - \bar{x})(x_t - \bar{x})^*. \tag{26.2}$$

Из усиленного закона больших чисел вытекает, что оценки

$$\bar{x} = \frac{1}{N} \sum_{t=1}^{N} x_t, \qquad S = \frac{1}{N-1} \sum_{t=2}^{N} z_t z_t^*$$

сильно состоятельны. Оценка $\hat{\Sigma}$ отличается от S только множителем $(N-1)/N$, который сходится к единице. Значит, она тоже сильно состоятельна.

Задача. Проверьте, что случайные величины $\bar{x}$ и $x_t - \bar{x}$ некоррелированы, и выведите отсюда независимость оценок $\bar{x}$ и $\hat{\Sigma}$.

Теорема 26.3. *Элементы матриц $\hat{\Sigma} = (\hat{\sigma}_{ij})_{i,j=1}^n$ и $S = (s_{ij})_{i,j=1}^n$ имеют совместные асимптотически нормальные распределения, причем*

$$\lim_{N\to\infty} N \operatorname{Cov}\{\hat{\sigma}_{ij}, \hat{\sigma}_{kl}\} = (N-1)\operatorname{Cov}\{s_{ij}, s_{kl}\} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{kj}.$$

Доказательство. В силу центральной предельной теоремы совокупность всех элементов матрицы S асимптотически нормально распределена. Найдем их ковариации с помощью формул (26.2) и (26.1):

$$(N-1)\operatorname{Cov}\{s_{ij}, s_{kl}\} = \frac{1}{N-1}\operatorname{Cov}\{N\hat{\sigma}_{ij}, N\hat{\sigma}_{kl}\} =$$

$$= \frac{1}{N-1}\mathsf{E}\Big\{\sum_{t=2}^{N}\sum_{s=2}^{N}\big(z_{ti}z_{tj} - \mathsf{E}\{z_{ti}z_{tj}\}\big)\big(z_{sk}z_{sl} - \mathsf{E}\{z_{sk}z_{sl}\}\big)\Big\} =$$

$$= \mathsf{E}\Big\{\big(z_{ti}z_{tj} - \mathsf{E}\{z_{ti}z_{tj}\}\big)\big(z_{tk}z_{tl} - \mathsf{E}\{z_{tk}z_{tl}\}\big)\Big\} =$$

$$= \mathsf{E}\{z_{ti}z_{tj}z_{tk}z_{tl}\} - \mathsf{E}\{z_{ti}z_{tj}\}\mathsf{E}\{z_{tk}z_{tl}\} =$$

$$= \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{kj} - \sigma_{ij}\sigma_{kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{kj},$$

где в последней строке мы использовали равенство (22.2) из предложения 22.1. □

§ 27. Оценки максимального правдоподобия для параметров нормальных распределений

В следующей теореме доказывается, что в случае семейства нормальных распределений выборочное среднее $\bar{x}$ и выборочная матрица ковариаций $\hat{\Sigma}$ будут оценками максимального правдоподобия.

Теорема 27.1. *Пусть $X = (x_1, \ldots, x_N)$ — выборка из нормального распределения $\mathcal{N}_n(\mu, \Sigma)$. Если $N > n$ и $|\Sigma| \neq 0$, то с вероятностью 1 статистики $\bar{x}$ и $\hat{\Sigma}$ будут оценками максимального правдоподобия.*

Доказательство. Плотность распределения выборки X имеет вид

$$p(X) = \prod_{t=1}^{N} p(x_t) = \prod_{t=1}^{N} \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\Big\{-\frac{1}{2}(x_t - \mu)^* \Sigma^{-1}(x_t - \mu)\Big\}.$$

Ей отвечает логарифмическая функция правдоподобия

$$l(\mu,\Sigma)=\ln p(X)=-\frac{nN}{2}\ln(2\pi)+\frac{N}{2}\ln\left|\Sigma^{-1}\right|-\frac{1}{2}\sum_{t=1}^{N}(x_t-\mu)^*\Sigma^{-1}(x_t-\mu). \tag{27.1}$$

Напомним, что оценка максимального правдоподобия — это точка максимума функции $l(\mu,\Sigma)$, где переменная Σ является симметричной положительно определенной матрицей. Поэтому утверждение теоремы вытекает из следующей леммы.

Лемма 27.2. *Пусть матрица Σ симметрична и положительно определена. Тогда функция $l(\mu,\Sigma)$ из (27.1) достигает максимума по переменной μ при $\mu=\bar{x}$, по переменной Σ при*

$$\Sigma=\frac{1}{N}\sum_{t=1}^{N}(x_t-\mu)(x_t-\mu)^*,$$

а по совокупности переменных при $\mu=\bar{x}$ и $\Sigma=\hat{\Sigma}$.

Доказательство. Максимум $l(\mu,\Sigma)$ по μ достигается в той же точке, что и минимум функции $\sum_{t=1}^{N}(x_t-\mu)^*\Sigma^{-1}(x_t-\mu)$. Заметим, что $\sum_{t=1}^{N}(x_t-\bar{x})=0$. Поэтому

$$\sum_{t=1}^{N}(x_t-\mu)^*\Sigma^{-1}(x_t-\mu)=$$

$$=\sum_{t=1}^{N}\big((x_t-\bar{x})+(\bar{x}-\mu)\big)^*\Sigma^{-1}\big((x_t-\bar{x})+(\bar{x}-\mu)\big)=$$

$$=\sum_{t=1}^{N}(x_t-\bar{x})^*\Sigma^{-1}(x_t-\bar{x})+N(\bar{x}-\mu)^*\Sigma^{-1}(\bar{x}-\mu).$$

Поскольку матрица Σ^{-1}, как и Σ, симметрична и положительно определена, последнее выражение минимально при $\mu=\bar{x}$.

Теперь найдем максимум функции $l(\mu,\Sigma)$ по Σ. Будем считать, что

$$x_t=(x_{t1},\ldots,x_{tn}),\qquad \mu=(\mu_1,\ldots,\mu_n),$$

$$\Sigma=\big(\sigma_{ij}\big)_{i,j=1}^{n},\qquad \Sigma^{-1}=\big(\sigma^{ij}\big)_{i,j=1}^{n}.$$

Если матрицы Σ и Σ^{-1} невырождены, то они взаимно однозначно связаны друг с другом. Нам будет удобно искать максимум у $l(\mu, \Sigma)$ не по отношению к Σ, а по отношению к Σ^{-1}. Необходимым условием такого максимума является обращение в нуль всех частных производных от функции $l(\mu, \Sigma)$ по переменным σ^{ij}. Прежде чем вычислять эти производные, выпишем несколько вспомогательных равенств.

Обозначим через b_{ij} алгебраические дополнения элементов σ^{ij} в матрице Σ^{-1}. Из линейной алгебры известны формула разложения определителя по строке

$$\left|\Sigma^{-1}\right| = \sum_{k=1}^{n} \sigma^{ik} b_{ik}, \qquad i = 1, \ldots, n,$$

и формула Крамера для элементов обратной матрицы

$$\sigma_{ji} = \frac{b_{ij}}{|\Sigma^{-1}|}.$$

Очевидно,

$$\sum_{t=1}^{N} (x_t - \mu)^* \Sigma^{-1} (x_t - \mu) = \sum_{t=1}^{N} \sum_{i,j=1}^{n} (x_{ti} - \mu_i) \sigma^{ij} (x_{tj} - \mu_j).$$

Для сокращения записи введем еще матрицу $A = (a_{ij})_{i,j=1}^{n}$:

$$A = \sum_{t=1}^{N} (x_t - \mu)(x_t - \mu)^*, \qquad a_{ij} = \sum_{t=1}^{N} (x_{ti} - \mu_i)(x_{tj} - \mu_j).$$

Используя эти равенства и формулу (27.1), находим производные

$$\frac{\partial l(\mu, \Sigma)}{\partial \sigma^{ij}} = \frac{N}{2} \frac{b_{ij}}{|\Sigma^{-1}|} - \frac{1}{2} \sum_{t=1}^{N} (x_{ti} - \mu_i)(x_{tj} - \mu_j) = \frac{1}{2} (N \sigma_{ji} - a_{ij}).$$

Они обращаются в нуль лишь при $\sigma_{ji} = \frac{1}{N} a_{ij}$ или, в матричной записи, при $\Sigma = \frac{1}{N} A$. Тем самым лемма доказана. $\square$

Задача. На самом деле мы доказали теорему в предположении невырожденности матрицы A. Проверьте, что при $N > n$ это условие выполняется с вероятностью 1.

Как и всякие оценки максимального правдоподобия, выборочное среднее $\bar{x}$ и выборочная ковариационная матрица $\hat{\Sigma}$ обладают свойствами сильной состоятельности, асимптотической несмещенности и асимптотической нормальности. Но мы это уже доказали независимо в предыдущем параграфе.

§ 28. Выборочные корреляции†

Напомним определение коэффициента корреляции между случайными величинами x и y:

$$r_{xy} = \operatorname{Corr}\{x, y\} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}.$$

В этой формуле σ_{xy} обозначает ковариацию

$$\sigma_{xy} = \operatorname{Cov}\{x, y\} = \mathsf{E}\big\{(x - \mathsf{E}x)(y - \mathsf{E}y)\big\},$$

а ковариации σ_{xx} и σ_{yy} определяются аналогично.

Пусть заданы две выборки $X = (x_1, \ldots, x_N)$ и $Y = (y_1, \ldots, y_N)$. Тогда выборочная оценка для r_{xy} имеет вид

$$\hat{r}_{xy} = \frac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_{xx}\hat{\sigma}_{yy}}}, \qquad \text{где} \ \ \hat{\sigma}_{xy} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y}),$$

и величины $\hat{\sigma}_{xx}$, $\hat{\sigma}_{yy}$ определяются аналогично. Сейчас наша цель — вычислить асимптотику для распределения статистики $\hat{r}_{xy}$.

Лемма 28.1. *Пусть функция $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ дифференцируема в точке $a \in \mathbb{R}^n$, и последовательность случайных величин $u_N \in \mathbb{R}^n$ такова, что распределение величины $\sqrt{N}(u_N - a)$ сходится к $\mathcal{N}_n(0, \Sigma)$. Тогда распределение выражения $\sqrt{N}\big(\varphi(u_N) - \varphi(a)\big)$ сходится к нормальному закону $\mathcal{N}_1\big(0,\, \nabla^*\varphi(a)\Sigma\nabla\varphi(a)\big)$, где $\nabla\varphi(a) = \operatorname{grad}\varphi(a)$.*

Доказательство. По определению дифференцируемости

$$\begin{aligned} \sqrt{N}\big(\varphi(u_N) - \varphi(a)\big) &= \sqrt{N}\,\nabla^*\varphi(a)(u_N - a) + \sqrt{N}o(u_N - a) = \\ &= \nabla^*\varphi(a)\sqrt{N}(u_N - a) + \alpha(u_N - a)\sqrt{N}\,|u_N - a|, \end{aligned} \tag{28.1}$$

где $\alpha(v)$ — бесконечно малая функция при $v \to 0$. Из предложения 22.2 следует, что распределение случайной величины $\nabla^*\varphi(a)\sqrt{N}(u_N - a)$ сходится к нормальному закону $\mathcal{N}_1\big(0,\, \nabla^*\varphi(a)\Sigma\nabla\varphi(a)\big)$. А случайная величина $\alpha(u_N-a)\sqrt{N}\,|u_N-a|$ сходится по вероятности к нулю.

Конечно, этих рассуждений еще не достаточно для строгого доказательства. Для полной строгости нужно еще проверить, что величина $\alpha(u_N-a)\sqrt{N}\,|u_N-a|$ действительно сходится к нулю по вероятности, дать определение сходимости многомерных распределений, проверить его для распределения суммы (28.1). Мы опускаем все эти технические детали. □

Теорема 28.2. *Если пара случайных величин x, y имеет совместное нормальное распределение и $|r_{xy}| \neq 1$, то распределение случайной величины $\sqrt{N}(\hat{r}_{xy}-r_{xy})$ сходится к $\mathcal{N}_1\big(0,\,(1-r_{xy}^2)^2\big)$.*

Доказательство. Рассмотрим последовательность трехмерных векторов

$$u_N = (u_{N1}, u_{N2}, u_{N3}) = \left(\frac{\hat{\sigma}_{xx}}{\sigma_{xx}},\ \frac{\hat{\sigma}_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}},\ \frac{\hat{\sigma}_{yy}}{\sigma_{yy}}\right).$$

Поскольку выборочные ковариации сильно состоятельны, u_N почти наверное сходится к вектору $a = (1, r_{xy}, 1)$. С помощью теоремы 26.3 нетрудно доказать, что распределение вектора $\sqrt{N}(u_N-a)$ сходится к $\mathcal{N}_3(0,\Sigma_3)$, где

$$\Sigma_3 = \begin{pmatrix} 2 & 2r_{xy} & 2r_{xy}^2 \\ 2r_{xy} & 1+r_{xy}^2 & 2r_{xy} \\ 2r_{xy}^2 & 2r_{xy} & 2 \end{pmatrix}.$$

Выразим выборочную корреляцию $\hat{r}_{xy}$ через координаты вектора u_N:

$$\hat{r}_{xy} = \varphi(u_N) = \frac{u_{N2}}{\sqrt{u_{N1}u_{N3}}}.$$

Вычисляя частные производные функции $\varphi(u_N)$ в точке a, получаем равенство

$$\nabla\varphi(a) = \Big(-\frac{r_{xy}}{2},\ 1,\ -\frac{r_{xy}}{2}\Big).$$

По лемме 28.1 распределение величины $\sqrt{N}(\hat{r}_{xy}-r_{xy}) = \sqrt{N}(\varphi(u_N)-\varphi(a))$ сходится к нормальному закону $\mathcal{N}_1\big(0,\,\nabla^*\varphi(a)\Sigma_3\nabla\varphi(a)\big)$. Наконец, прямая выкладка показывает, что значение выражения $\nabla^*\varphi(a)\Sigma_3\nabla\varphi(a)$ равно $(1-r_{xy}^2)^2$. □

Следствие 28.3. *Если функция $\psi : \mathbb{R} \to \mathbb{R}$ дифференцируема в точке r_{xy}, то в условиях теоремы 28.2 распределение случайной величины $\sqrt{N}\big(\psi(\hat{r}_{xy})-\psi(r_{xy})\big)$ сходится к $\mathcal{N}_1\big(0,\,\psi'(r_{xy})^2(1-r_{xy}^2)^2\big)$.*

Доказательство. Это непосредственно вытекает из леммы 28.1 и теоремы 28.2. □

Подберем такую функцию ψ, для которой выполнялось бы равенство $\psi'(u)^2(1-u^2)^2 = 1$. Для этого достаточно решить простейшее дифференциальное уравнение

$$\psi'(u) = \frac{1}{1-u^2}.$$

Интегрируя его, получаем $\psi(u) = \frac{1}{2}\ln\frac{1+u}{1-u}$. В прикладной статистике эта функция называется Z-преобразованием Фишера, а функция

$$Z(\hat{r}_{xy}) = \frac{1}{2}\ln\frac{1+\hat{r}_{xy}}{1-\hat{r}_{xy}}$$

называется Z*-статистикой Фишера.* По следствию 28.3 распределение случайной величины $\sqrt{N}(Z(\hat{r}_{xy})-Z(r_{xy}))$ сходится к стандартному нормальному закону $\mathcal{N}_1(0,1)$.

В частности, для любого $\varepsilon \in (0,1)$ и квантили $\Delta = \Phi^{-1}(1-\varepsilon/2)$ имеет место асимптотика

$$P\Big\{\sqrt{N}\,\big|Z(\hat{r}_{xy}) - Z(r_{xy})\big| \leqslant \Delta\Big\} \longrightarrow 1-\varepsilon. \tag{28.2}$$

С помощью этой асимптотики можно построить решающее правило для проверки гипотезы о значении коэффициента корреляции. Пусть задано некоторое число $r \in (-1,1)$, и гипотеза H_0 состоит в том, что $r_{xy} = r$, а гипотеза H_1 состоит в том, что $r_{xy} \neq r$. Тогда при большом объеме выборки N решающее правило имеет вид

$$\begin{cases} r_{xy} = r, & \text{если}\ \ \sqrt{N}\,\big|Z(\hat{r}_{xy}) - Z(r)\big| \leqslant \Delta, \\ r_{xy} \neq r, & \text{если}\ \ \sqrt{N}\,\big|Z(\hat{r}_{xy}) - Z(r)\big| > \Delta, \end{cases} \tag{28.3}$$

где число Δ определяется по уровню значимости ε как квантиль $\Delta = \Phi^{-1}(1-\varepsilon/2)$. Часто в данном критерии вместо $\sqrt{N}$ пишут $\sqrt{N-1}$; тогда он оказывается немного (на самом деле незначительно) точнее.

Если $r = 0$, то гипотеза $r_{xy} = 0$ называется гипотезой независимости случайных величин x и y (потому что в ситуации, когда они нормально распределены, их независимость равносильна некоррелированности). Однако для проверки гипотезы независимости, особенно при небольших объемах выборки, рекомендуется использовать более точный критерий, который будет описан в § 33 (см. формулу (33.7)).

Асимптотика (28.2) позволяет построить доверительный интервал для r_{xy} с доверительной вероятностью $1-\varepsilon$. Он имеет вид

$$\left(\operatorname{th}\Big\{Z(\hat{r}_{xy}) - \frac{\Delta}{\sqrt{N}}\Big\},\ \operatorname{th}\Big\{Z(\hat{r}_{xy}) + \frac{\Delta}{\sqrt{N}}\Big\}\right),$$

где $\operatorname{th}(v) = \frac{e^v-e^{-v}}{e^v+e^{-v}}$ — функция, обратная к преобразованию Фишера.

Глава 4. ЛИНЕЙНАЯ РЕГРЕССИЯ

§ 29. Линейная регрессия случайных величин

В самом общем смысле линейная регрессия — это наука о том, как наилучшим образом приблизить произвольный вектор из евклидова или гильбертова пространства вектором из некоторого заранее заданного конечномерного подпространства. Разумеется, таким наилучшим приближением будет ортогональная проекция исходного вектора на указанное подпространство. Вначале мы рассмотрим эту проекцию в ситуации, когда гильбертово пространство состоит из случайных величин.

Пусть $(\Omega, \mathfrak{A}, P)$ — некоторое вероятностное пространство. Обозначим буквой H гильбертово пространство вещественнозначных случайных величин $L^2(\Omega, \mathfrak{A}, P)$ со скалярным произведением

$$(x, y) = \int_\Omega x(\omega) y(\omega)\, dP(\omega) = \mathsf{E}\{xy\}.$$

Рассмотрим конечный набор случайных величин $x = (x_1, \ldots, x_n)$, где $x_i \in H$. Обозначим через $L(x) = \ell(1, x_1, \ldots, x_n)$ линейную оболочку функций 1, x_1, $\ldots$, x_n:

$$L(x) = \{\alpha_0 + \alpha_1 x_1 + \ \ldots\ + \alpha_n x_n \mid \alpha_i \in \mathbb{R}\}.$$

Линейной регрессией любой случайной величины $y \in H$ на совокупность $x = (x_1, \ldots, x_n)$ называется ортогональная проекция y на подпространство $L(x)$. Будем обозначать эту регрессию как $P_x y$.

Ортогональная проекция вектора y на линейное подпространство $L(x)$ является ближайшей к y точкой из $L(x)$. Запишем этот факт в виде следующего предложения.

Предложение 29.1. *Если $P_x y$ — линейная регрессия случайной величины y на совокупность $x = (x_1, \ldots, x_n)$, то для любой другой случайной величины $z \in L(x)$ выполняется неравенство*

$$\mathsf{E}\big\{(y - P_x y)^2\big\} \leqslant \mathsf{E}\big\{(y - z)^2\big\}.$$

Оно обращается в равенство только в том случае, когда $z = P_x y$ п. н.

Вычислим линейную регрессию y на x в явном виде. По определению, вектор $y - P_x y$ должен быть ортогонален подпространству $L(x)$. Условия этой ортогональности можно записать так:

$$\mathsf{E}\{(y - P_x y) \cdot 1\} = 0, \qquad \mathsf{E}\{(y - P_x y) x_i\} = 0, \quad i = 1, \ldots, n.$$

Они равносильны следующим двум равенствам:

$$\mathsf{E}\{y - P_x y\} = 0, \qquad \mathrm{Cov}\{y - P_x y, x\} = 0. \tag{29.1}$$

Введем обозначения

$$\Sigma_{xx} = \mathrm{Cov}\{x, x\} = \mathsf{E}\big\{(x - \mathsf{E}x)(x - \mathsf{E}x)^*\big\},$$
$$\Sigma_{yx} = \mathrm{Cov}\{y, x\} = \mathsf{E}\big\{(y - \mathsf{E}y)(x - \mathsf{E}x)^*\big\}$$

(здесь Σ_{xx} — матрица $n \times n$, а Σ_{yx} — вектор-строка размерности n).

Предложение 29.2. *Если матрица Σ_{xx} невырождена, то тогда линейная регрессия y на x имеет вид*

$$P_x y = \mathsf{E}y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mathsf{E}x). \tag{29.2}$$

Доказательство. Регрессию $P_x y \in L(x)$ можно представить как линейную комбинацию

$$P_x y = \alpha_0 + \sum_{i=1}^{n} \alpha_i (x_i - \mathsf{E}x_i) = \alpha_0 + \alpha^* (x - \mathsf{E}x), \tag{29.3}$$

где α^* — вектор-строка $(\alpha_1, \ldots, \alpha_n)$. В силу левого равенства (29.1) из (29.3) следует, что $\alpha_0 = \mathsf{E}\{P_x y\} = \mathsf{E}y$. Из правого равенства (29.1) вытекает, что

$$\Sigma_{yx} = \mathrm{Cov}\{y, x\} = \mathrm{Cov}\{P_x y, x\} = \mathsf{E}\big\{\alpha^* (x - \mathsf{E}x)(x - \mathsf{E}x)^*\big\} = \alpha^* \Sigma_{xx}.$$

Следовательно, $\alpha^* = \Sigma_{yx} \Sigma_{xx}^{-1}$. Подставляя найденные α_0 и α^* в (29.3), получаем (29.2). $\square$

Линейная регрессия случайного вектора $y = (y_1, \ldots, y_m)$ на совокупность случайных величин $x = (x_1, \ldots, x_n)$ определяется покомпонентно: $P_x y = (P_x y_1, \ldots, P_x y_m)$. Мы доказали, что

$$P_x y_i = \mathsf{E}y_i + \Sigma_{y_i x} \Sigma_{xx}^{-1} (x - \mathsf{E}x).$$

Объединяя эти равенства для $i = 1, \ldots, n$, мы получим ту же самую формулу (29.2), но только в ней $P_x y$ и $\mathsf{E} y$ будут вектор-столбцами размерности m, а Σ_{yx} — матрицей размерности $m \times n$.

Предложение 29.3. *Если $P_x y$ — линейная регрессия случайного вектора $y = (y_1, \ldots, y_m)$ на совокупность $x = (x_1, \ldots, x_n)$ и $|\Sigma_{xx}| \neq 0$, то*

$$\operatorname{Cov}\{y - P_x y, P_x y\} = 0, \tag{29.4}$$

$$\operatorname{Cov}\{P_x y, P_x y\} = \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}, \tag{29.5}$$

$$\operatorname{Cov}\{y - P_x y, y - P_x y\} = \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}. \tag{29.6}$$

Доказательство. Случайные величины $P_x y_i$ лежат в линейном подпространстве $L(x) = \ell(1, x_1, \ldots, x_n)$, а разности $y_i - P_x y_i$ ортогональны ему. Значит,

$$\operatorname{Cov}\{y_i - P_x y_i, P_x y_j\} = \mathsf{E}\big\{(y_i - P_x y_i)(P_x y_j - \mathsf{E}\{P_x y_j\})\big\} = 0$$

при всех i, j, что равносильно одному матричному равенству (29.4). Для вычисления матрицы ковариаций $\operatorname{Cov}\{P_x y, P_x y\}$ подставим в нее формулу (29.2):

$$\operatorname{Cov}\{P_x y, P_x y\} = \mathsf{E}\big\{\Sigma_{yx} \Sigma_{xx}^{-1} (x - \mathsf{E} x)(x - \mathsf{E} x)^* \Sigma_{xx}^{-1} \Sigma_{xy}\big\} =$$

$$= \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xx} \Sigma_{xx}^{-1} \Sigma_{xy} = \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}.$$

Наконец, (29.6) получается из (29.4) и (29.5), если раскрыть выражение $\operatorname{Cov}\{y - P_x y, y - P_x y\}$ по линейности. □

Предложение 29.4. *Если два случайных вектора x, y имеют совместное нормальное распределение, то разность $y - P_x y$ нормально распределена и не зависит от x.*

Доказательство. Пара $(x, y - P_x y)$ линейно зависит от x и y. Значит, она имеет совместное нормальное распределение. По определению регрессии случайные векторы x и $y - P_x y$ некоррелированы (см. правое равенство (29.1)) и, следовательно, независимы. □

§ 30. Множественная и частная корреляции

Множественной корреляцией между случайной величиной $y \in H$ и набором случайных величин $x = (x_1, \ldots, x_n)$ называется корреляция между y и линейной регрессией y на x: $R_{yx} = \operatorname{Corr}\{y, P_x y\}$.

Напомним, что величина $\mathrm{Corr}\{y, P_x y\}$ равна косинусу угла между векторами $y - \mathsf{E}y$ и $P_x y - \mathsf{E}y$ в гильбертовом пространстве случайных величин (мы здесь пользуемся тем, что $\mathsf{E}\{P_x y\} = \mathsf{E}y$). По определению, регрессия $P_x y$ является ортогональной проекцией y на линейное подпространство $L(x)$. Это подпространство содержит константу $\mathsf{E}y$. Поэтому ортогональной проекцией разности $y - \mathsf{E}y$ будет $P_x y - \mathsf{E}y$. По построению эти два вектора являются гипотенузой и катетом прямоугольного треугольника, а косинус угла между ними равен отношению длины катета к длине гипотенузы

$$R_{yx} = \frac{\sqrt{\mathsf{E}\big\{(P_x y - \mathsf{E}y)^2\big\}}}{\sqrt{\mathsf{E}\big\{(y - \mathsf{E}y)^2\big\}}} = \frac{\sqrt{\mathsf{D}\{P_x y\}}}{\sqrt{\mathsf{D}y}}. \tag{30.1}$$

Используя равенства $\mathsf{D}\{P_x y\} = \mathrm{Cov}\{P_x y, P_x y\}$ и (29.5), получаем, что

$$R_{yx} = \sqrt{\frac{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}}{\Sigma_{yy}}}. \tag{30.2}$$

В частности, если случайная величина x одномерна, то тогда из (30.2) вытекает равенство $R_{yx} = |\mathrm{Corr}\{y, x\}|$.

Множественную корреляцию R_{yx} часто интерпретируют как меру близости между случайной величиной y и ее регрессией $P_x y$ (или как меру количества информации об y, содержащейся в x). Чем ближе эта корреляция к единице, тем лучше регрессия приближает исходную случайную величину (и тем точнее можно определить y по известному x). При единичной корреляции имеет место равенство $y = P_x y$ (и тогда y однозначно определяется по x), а при нулевой корреляции считается, что знание x не дает никакой информации об y.

Пусть заданы две случайные величины y, z и случайный вектор $x = (x_1, \ldots, x_n)$. Тогда *частной дисперсией* случайной величины y при фиксированном x называется число $\mathsf{D}\{y|x\} = \mathsf{D}\{y - P_x y\}$, *частной ковариацией* случайных величин y и z при фиксированном x называется число $\mathrm{Cov}\{y, z|x\} = \mathrm{Cov}\{y - P_x y, z - P_x z\}$, а *частной корреляцией* между y и z называется число $R_{yz|x} = \mathrm{Corr}\{y - P_x y, z - P_x z\}$. Из этих определений следует, что

$$R_{yz|x} = \frac{\mathrm{Cov}\{y, z|x\}}{\sqrt{\mathsf{D}\{y|x\}\mathsf{D}\{z|x\}}}. \tag{30.3}$$

По определению регрессии вектор $y - P_x y$ ортогонален линейному подпространству $L(x)$, а вектор $P_x y - \mathsf{E}y$ принадлежит ему. Значит, эти два вектора ортогональны и являются катетами прямоугольного треугольника с гипотенузой $y - \mathsf{E}y$. Из теоремы Пифагора и (30.1) вытекает равенство

$$\mathsf{D}y = \mathsf{D}\{y - P_x y\} + \mathsf{D}\{P_x y\} = \mathsf{D}\{y|x\} + R^2_{yx}\mathsf{D}y. \tag{30.4}$$

Пусть имеются два набора случайных величин $y = (y_1, \ldots, y_m)$ и $x = (x_1, \ldots, x_n)$. Матрицей частных ковариаций случайного вектора y при фиксированном x называется матрица

$$\Sigma_{yy|x} = \mathrm{Cov}\{y - P_x y, y - P_x y\}.$$

Ее элементы

$$\Sigma_{y_i y_j|x} = \mathrm{Cov}\{y_i - P_x y_i, y_j - P_x y_j\}$$

являются частными ковариациями компонент y_i и y_j, а диагональные элементы $\Sigma_{y_i y_i|x}$ совпадают с $\mathsf{D}\{y_i|x\}$. Из (29.6) следует, что

$$\Sigma_{yy|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}, \tag{30.5}$$

$$\Sigma_{y_i y_j|x} = \Sigma_{y_i y_j} - \Sigma_{y_i x}\Sigma_{xx}^{-1}\Sigma_{x y_j}. \tag{30.6}$$

А из равенства (30.3) следует, что частные корреляции между y_i и y_j вычисляются по элементам матрицы частных ковариаций $\Sigma_{yy|x}$ так:

$$R_{y_i y_j|x} = \frac{\Sigma_{y_i y_j|x}}{\sqrt{\Sigma_{y_i y_i|x}\Sigma_{y_j y_j|x}}}. \tag{30.7}$$

Теорема 30.1. *Если два случайных вектора x, y имеют совместное нормальное распределение, то линейная регрессия $P_x y$ совпадает с условным математическим ожиданием $\mathsf{E}\{y|x\}$, а матрица частных ковариаций $\Sigma_{yy|x}$ совпадает с матрицей ковариаций условного распределения случайного вектора y при фиксированном x.*

Доказательство. В теореме 24.1 было доказано, что условное распределение случайного вектора y при фиксированном x является нормальным распределением с математическим ожиданием

$$\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)$$

и матрицей ковариаций

$$\Sigma_{yy|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy},$$

где используются обозначения $\mu_{y|x} = \mathsf{E}\{y|x\}$, $\mu_y = \mathsf{E}y$, $\mu_x = \mathsf{E}x$. С другой стороны, формулы (29.2) и (30.5) для линейной регрессии $P_x y$ и матрицы частных ковариаций $\Sigma_{yy|x}$ выглядят точно так же. □

§ 31. Многомерные выборочные оценки

Пусть $\Omega_N = \{1, \ldots, N\}$ — конечное вероятностное пространство с равномерным распределением вероятностей P (при котором вероятности элементарных событий 1, …, N одинаковы и равны $1/N$). Случайные величины на таком пространстве — это функции $X\colon \Omega_N \to \mathbb{R}$. Они полностью определяются конечным набором всех своих значений $X(1)$, …, $X(N)$. Поэтому их можно представлять себе как N-мерные векторы $X = (X(1), \ldots, X(N))$. Гильбертово пространство случайных величин $L^2(\Omega_N, P)$ естественным образом отождествляется с евклидовым пространством $\mathbb{R}^N$, причем скалярные произведения в $L^2(\Omega_N, P)$ и в $\mathbb{R}^N$ отличаются лишь постоянным множителем N:

$$(X, Y)_{L^2(\Omega_N, P)} = \mathsf{E}\{XY\} = \frac{1}{N}\sum_{t=1}^{N} X(t)Y(t),$$

$$(X, Y)_{\mathbb{R}^N} = X^*Y = \sum_{t=1}^{N} X(t)Y(t).$$

Отсюда видно, что углы между векторами и ортогональные проекции в этих двух пространствах *совпадают.*

Напомним общий принцип построения выборочных оценок (см. § 3 из первой части этого курса). Всякая выборка $X = (x_1, \ldots, x_N) \in \mathbb{R}^N$ отождествляется со случайной величиной $X : \Omega_N \to \mathbb{R}$, сопоставляющей каждому номеру $t \in \Omega_N$ значение x_t. Если задано несколько выборок $X_j = (x_{1j}, \ldots, x_{Nj}) \in \mathbb{R}^N$, то любая их выборочная «характеристика» (где слово «характеристика» представляет собой лингвистическую переменную, принимающую значения «математическое ожидание», «дисперсия», «ковариация», «множественная корреляция» и т. п.) определяется как та же самая «характеристика» для соответствующих случайных величин $X_j\colon \Omega_N \to \mathbb{R}$ (со значениями $X_j(t) = x_{tj}$).

Приведем несколько примеров использования этого принципа.

1. Выборочное математическое ожидание (или выборочное среднее) для выборки $X = (x_1, \ldots, x_N)$ — это число

$$\bar{x} = \mathsf{E}X = \frac{1}{N}\sum_{t=1}^{N} x_t.$$

Часто бывает полезно воспринимать выборочное среднее не как число, а как постоянную функцию на пространстве Ω_N. Тогда эта функция отождествляется с постоянной выборкой $\bar{X} = (\bar{x}, \ldots, \bar{x}) \in \mathbb{R}^N$.

2. Выборочная дисперсия выборки $X = (x_1, \ldots, x_N)$ — это число

$$\hat{\sigma}^2 = \mathsf{E}\bigl\{(X - \bar{X})^2\bigr\} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2 = \frac{1}{N}\bigl|X - \bar{X}\bigr|^2.$$

3. Выборочный коэффициент корреляции между двумя выборками $X = (x_1, \ldots, x_N)$ и $Y = (y_1, \ldots, y_N)$ по определению равен косинусу угла между векторами $X - \bar{X}$ и $Y - \bar{Y}$ в гильбертовом пространстве $L^2(\Omega_N, P)$, или, что то же самое, в евклидовом пространстве $\mathbb{R}^N$:

$$\hat{r}_{xy} = \cos\bigl(X - \bar{X}, Y - \bar{Y}\bigr) = \frac{\bigl(X - \bar{X}, Y - \bar{Y}\bigr)}{\bigl|X - \bar{X}\bigr|\,\bigl|Y - \bar{Y}\bigr|}. \tag{31.1}$$

4. Предположим, что заданы два набора выборок $X_1, \ldots, X_n \in \mathbb{R}^N$ и $Y_1, \ldots, Y_m \in \mathbb{R}^N$, где $X_i = (x_{1i}, \ldots, x_{Ni})$ и $Y_j = (y_{1j}, \ldots, y_{Nj})$. Тогда выборочная матрица ковариаций $\hat{\Sigma}_{xy}$ размерности $n \times m$ состоит из элементов

$$\hat{\sigma}_{ij} = \frac{1}{N}\sum_{t=1}^{N}(x_{ti} - \bar{x}_i)(y_{tj} - \bar{y}_j) = \frac{1}{N}\bigl(X_i - \bar{X}_i, Y_j - \bar{Y}_j\bigr).$$

Объединяя эти равенства для всех i, j, можно записать их как одно матричное равенство

$$\hat{\Sigma}_{xy} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y})^*, \tag{31.2}$$

где $x_t = (x_{t1}, \ldots, x_{tn}) \in \mathbb{R}^n$ и $y_t = (y_{t1}, \ldots, y_{tm}) \in \mathbb{R}^m$. В частности, если $Y = X$, то

$$\hat{\Sigma}_{xx} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(x_t - \bar{x})^*. \tag{31.3}$$

5. По выборочной матрице ковариаций $\hat{\Sigma}_{xx} = (\hat{\sigma}_{ij})$ определяется выборочная корреляционная матрица $\hat{R}_{xx} = (\hat{r}_{ij})$ с элементами

$$\hat{r}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\,\hat{\sigma}_{jj}}} = \frac{\bigl(X_i - \bar{X}_i, X_j - \bar{X}_j\bigr)}{\bigl|X_i - \bar{X}_i\bigr|\,\bigl|X_j - \bar{X}_j\bigr|}. \tag{31.4}$$

6. Пусть заданы выборки $X_1, \ldots, X_n \in \mathbb{R}^N$ и $Y \in \mathbb{R}^N$. Определим еще единичную выборку $X_0 = (1, \ldots, 1) \in \mathbb{R}^N$. Обозначим через $L(X) = \ell(X_0, X_1, \ldots, X_n)$ линейную оболочку векторов X_0, X_1, …, X_n в пространстве $\mathbb{R}^N$, а через P_X — ортогональную проекцию всего $\mathbb{R}^N$ на подпространство $L(X)$. Тогда проекция разности $Y - \bar{Y}$ на $L(X)$ будет равна $P_X Y - \bar{Y}$ (поскольку вектор $\bar{Y}$ коллинеарен с $X_0 \in L(X)$).

Выборочная множественная корреляция $\hat{R}_{yx}$ — это косинус угла между вектором $Y - \bar{Y}$ и его проекцией $P_X Y - \bar{Y}$ на $L(X)$. Поскольку векторы $Y - \bar{Y}$ и $P_X Y - \bar{Y}$ являются соответственно гипотенузой и катетом прямоугольного треугольника,

$$\hat{R}_{yx} = \cos\bigl(Y - \bar{Y}, P_X Y - \bar{Y}\bigr) = \frac{\bigl|P_X Y - \bar{Y}\bigr|}{\bigl|Y - \bar{Y}\bigr|}. \tag{31.5}$$

Альтернативное определение выборочной множественной корреляции можно получить из формулы (30.2), если заменить в ней все матрицы ковариаций на их выборочные оценки:

$$\hat{R}_{yx} = \sqrt{\frac{\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}}{\hat{\Sigma}_{yy}}}. \tag{31.6}$$

Подчеркнем, что формулы (31.5) и (31.6) дают одинаковый результат, потому что определяют *настоящую* множественную корреляцию для случайных величин на Ω_N, порождаемых выборками Y и X_i.

7. Предположим, что есть выборки $X_1, \ldots, X_n \in \mathbb{R}^N$ и $Y, Z \in \mathbb{R}^N$. Выборочной частной корреляцией между Y и Z при фиксированных X_1, …, X_n называется выборочная корреляция (то есть косинус угла) между $Y - P_X Y$ и $Z - P_X Z$ (где, как и раньше, P_X — ортогональная проекция всего $\mathbb{R}^N$ на $L(X) = \ell(X_0, X_1, \ldots, X_n)$). Другими словами,

$$\hat{R}_{yz|x} = \cos(Y - P_X Y, Z - P_X Z) = \frac{(Y - P_X Y, Z - P_X Z)}{|Y - P_X Y|\,|Z - P_X Z|}. \tag{31.7}$$

Эта формула полностью аналогична (30.3). А формулы (30.5) – (30.7) в рассматриваемой ситуации приобретают вид

$$\hat{R}_{yz|x} = \frac{\hat{\Sigma}_{yz|x}}{\sqrt{\hat{\Sigma}_{yy|x}\,\hat{\Sigma}_{zz|x}}}, \tag{31.8}$$

$$\hat{\Sigma}_{uv|x} = \hat{\Sigma}_{uv} - \hat{\Sigma}_{ux}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xv}, \tag{31.9}$$

где буквы u, v обозначают любые из переменных y, z.

§ 32. Метод наименьших квадратов

Пусть задан набор выборок $X_1, \ldots, X_n$ одинаковой мощности N, где $X_i = (x_{1i}, \ldots, x_{Ni})$, и еще одна выборка $Y = (y_1, \ldots, y_N)$. Можно представлять себе, что в моменты времени $t = 1, \ldots, N$ наблюдаются значения $x_{t1}, \ldots, x_{tn}$ и y_t. Предполагается, что значения y_t каким-то образом зависят от $x_{t1}, \ldots, x_{tn}$. Общая задача нахождения регрессии выборки Y на набор выборок $X_1, \ldots, X_n$ состоит в нахождении *функции регрессии* $y = f(x_1, \ldots, x_n)$, наилучшим образом отражающей эту зависимость. Обычно от функции регрессии требуют, чтобы совокупность разностей $y_t - f(x_{t1}, \ldots, x_{tn})$ при всех $t = 1, \ldots, N$ была бы в некотором смысле минимальной.

Часто мерой погрешности регрессии считают функционал

$$RSS(f) = \sum_{t=1}^{N} \bigl(y_t - f(x_{t1}, \ldots, x_{tn})\bigr)^2,$$

Он называется *остаточной суммой квадратов*. Метод наименьших квадратов заключается в нахождении функции f, минимизирующей этот функционал.

Поскольку самые простые функции — это линейные, естественно вначале попытаться найти регрессию в классе линейных функций.

Любая неоднородная линейная функция $f(x_1, \ldots, x_n)$ имеет вид

$$f(x) = \theta_0 + \sum_{i=1}^{n} \theta_i x_i.$$

Она полностью определяется набором параметров $\theta = (\theta_0, \theta_1, \ldots, \theta_n)$.

Таким образом, задача состоит в нахождении вектора θ, минимизирующего функционал

$$RSS(\theta) = \sum_{t=1}^{N}\Bigl(y_t - \theta_0 - \sum_{i=1}^{n}\theta_i x_{ti}\Bigr)^2 = \sum_{t=1}^{N}\Bigl(y_t - \sum_{i=0}^{n}\theta_i x_{ti}\Bigr)^2, \qquad (32.1)$$

где мы для единообразия записи полагаем $x_{t0} = 1$.

Определим единичную выборку $X_0 = (1, \ldots, 1) \in \mathbb{R}^N$. Составим из столбцов X_0, X_1, …, X_n матрицу $X = (x_{ti})$ размерности $N \times (n+1)$. Будем считать, что выборки X_i и Y являются элементами евклидова пространства $\mathbb{R}^N$. Тогда мы можем рассматривать остаточную сумму квадратов (32.1) как квадрат расстояния в $\mathbb{R}^N$ между вектором Y и линейной комбинацией $\theta_0 X_0 + \theta_1 X_1 + \ldots + \theta_n X_n = X\theta$:

$$RSS(\theta) = |Y - X\theta|^2. \qquad (32.2)$$

Множество всех линейных комбинаций векторов X_0, X_1, …, X_n образует линейное подпространство (линейную оболочку)

$$L(X) = \ell(X_0, X_1, \ldots, X_n) = \bigl\{X\theta \bigm| \theta \in \mathbb{R}^{n+1}\bigr\}.$$

А ближайшая к Y точка из подпространства $L(X)$ есть не что иное, как ортогональная проекция Y на $L(X)$.

Сейчас мы вычислим эту проекцию. Обозначим через X^* матрицу, сопряженную к X. Тогда произведение X^*X будет квадратной матрицей размерности $(n+1) \times (n+1)$.

Предложение 32.1. *Если для матрицы X размерности $N \times m$ выполняется условие $|X^*X| \neq 0$, то ортогональная проекция евклидова пространства $\mathbb{R}^N$ на подпространство $L(X) = \{X\theta \mid \theta \in \mathbb{R}^m\}$ задается матрицей $P_X = X(X^*X)^{-1}X^*$.*

Доказательство. Из определения пространства $L(X)$ следует, что для любого вектора $Y \in \mathbb{R}^N$ его образ $P_X Y = X(X^*X)^{-1}X^*Y$ принадлежит $L(X)$. Кроме того, для всякого вектора $X\theta \in L(X)$

$$(X\theta, Y - P_X Y) = \theta^* X^*(Y - P_X Y) = \theta^* X^* Y - \theta^* X^* X (X^*X)^{-1} X^* Y = 0.$$

Значит, вектор $Y - P_X Y$ ортогонален $L(X)$. □

В силу этого предложения ортогональная проекция вектора Y на подпространство $L(X)$ имеет вид

$$P_X Y = X\hat{\theta}, \quad \text{где} \quad \hat{\theta} = (X^*X)^{-1}X^*Y. \tag{32.3}$$

В математической статистике проекцию $P_X Y$ называют *множественной линейной регрессией* выборки Y на набор выборок $X_1, \ldots, X_n$.

Формула (32.3) перестает работать, если матрица X^*X вырождается. Это происходит тогда, когда размерность линейного пространства $L(X) = \ell(X_0, X_1, \ldots, X_n)$ оказывается меньше $n + 1$ (докажите это!). Если так случилось, то статистики говорят, что совокупность выборок $X_1, \ldots, X_n$ *мультиколлинеарна* (в отличие от всех остальных математиков, которые в той же самой ситуации сказали бы, что векторы $X_0, X_1, \ldots, X_n$ линейно зависимы). В случае мультиколлинеарности для нахождения линейной регрессии Y на X следует просто отбросить часть векторов $X_0, X_1, \ldots, X_n$, чтобы оставшиеся векторы образовывали базис подпространства $L(X)$.

Задача 1. Докажите равенство $P_X Y = \bar{Y} + \tilde{X}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}$, в котором матрица $\tilde{X}$ состоит из столбцов $X_1 - \bar{X}_1, \ldots, X_n - \bar{X}_n$.

Подсказка: транспонируйте формулу (29.2) или воспользуйтесь равенствами $N\hat{\Sigma}_{xx} = \tilde{X}^*\tilde{X}$ и $N\hat{\Sigma}_{xy} = \tilde{X}^*(Y - \bar{Y})$.

Задача 2*. Докажите равенство $|Y - P_X Y|^2 = |Z^*Z| \big/ |X^*X|$, где матрица Z состоит из столбцов $X_0, X_1, \ldots, X_n, Y$.

Подсказка: X^*X — матрица Грама для векторов $X_0, X_1, \ldots, X_n$, а определитель $|X^*X|$ совпадает с квадратом объема параллелепипеда, натянутого на них (см. приложение, п. II).

Задача 3. Докажите, что если $x_1, \ldots, x_n, y$ — элементы евклидова пространства, то ортогональной проекцией y на линейную оболочку $\ell(x_1, \ldots, x_n)$ будет вектор $\sum (y, x_i) g^{ij} x_j$, где g^{ij} — элементы матрицы, обратной к матрице Грама, состоящей из элементов $g_{ij} = (x_i, x_j)$.

В силу (32.3) остаточная сумма квадратов (32.2) становится минимальной при $\theta = \hat{\theta}$. Обозначим это минимальное значение через RSS:

$$RSS = \left| Y - X\hat{\theta} \right|^2 = |Y - P_X Y|^2 .$$

По теореме Пифагора $|Y - P_X Y|^2 + |P_X Y|^2 = |Y|^2$, откуда получаем, что

$$RSS = |Y|^2 - |P_X Y|^2 = Y^*Y - Y^*P_X Y.$$

Пусть $\bar{Y} = (\bar{y}, \ldots, \bar{y})$, где $\bar{y} = (y_1 + \ldots + y_N)/N$. Величина

$$TSS = \left|Y - \bar{Y}\right|^2 = \sum_{t=1}^{N} (y_t - \bar{y})^2$$

называется *полной суммой квадратов* для выборки Y, а величина

$$ESS = \left|P_X Y - \bar{Y}\right|^2$$

называется *объясненной суммой квадратов*. Считается, что ESS — это та часть вариации выборки Y, которую удается объяснить ее линейной зависимостью от X. По построению вектор $Y - P_X Y$ ортогонален линейному подпространству $L(X)$, а $P_X Y - \bar{Y}$ принадлежит ему. Значит, $|Y - P_X Y|^2 + |P_X Y - \bar{Y}|^2 = |Y - \bar{Y}|^2$, что равносильно тождеству

$$RSS + ESS = TSS.$$

Коэффициентом детерминации линейной регрессии называют число

$$R^2 = \frac{ESS}{TSS} = \frac{\left|P_X Y - \bar{Y}\right|^2}{\left|Y - \bar{Y}\right|^2}.$$

Оно совпадает с квадратом множественной корреляции $\hat{R}_{yx}$ (см. формулу (31.5)) и характеризует, насколько хорошо выборка Y приближается регрессией $P_X Y$.

Для каждой из трех сумм квадратов

$$TSS = \left|Y - \bar{Y}\right|^2, \quad RSS = \left|Y - P_X Y\right|^2, \quad ESS = \left|P_X Y - \bar{Y}\right|^2$$

определяется *число степеней свободы* как размерность линейного подпространства, содержащего соответствующий вектор $Y - \bar{Y}$, $Y - P_X Y$ или $P_X Y - \bar{Y}$. А именно, вектор $Y - \bar{Y}$ лежит в ортогональном дополнении к вектору X_0, имеющем размерность $N - 1$; вектор $Y - P_X Y$ лежит в ортогональном дополнении к $L(X)$, имеющем размерность $N - n - 1$ (при условии линейной независимости $X_0, X_1, \ldots, X_n$); а вектор $P_X Y - \bar{Y}$ лежит в ортогональном дополнении к X_0 в подпространстве $L(X)$, которое имеет размерность n. Отношения

$$\frac{TSS}{N-1}, \quad \frac{RSS}{N-n-1}, \quad \frac{ESS}{n}$$

называются соответственно *полным средним квадратом, остаточным средним квадратом* и *объясненным средним квадратом.*

§ 33. Модель множественной линейной регрессии

Моделью множественной линейной регрессии называется предположение о том, что некоторая выборка $Y = (y_1, \ldots, y_N)$ зависит от набора выборок $X_1, \ldots, X_n$, где $X_i = (x_{1i}, \ldots, x_{Ni})$, следующим образом:

$$y_t = \theta_0 + \theta_1 x_{t1} + \ldots + \theta_n x_{tn} + \xi_t, \qquad t = 1, \ldots, N, \tag{33.1}$$

где ξ_t — это последовательность независимых одинаково распределенных случайных величин с нулевым средним. Обычно ξ_t интерпретируются как погрешности измерения или случайные возмущения истинной функции регрессии

$$y = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n.$$

Выборочные значения x_{ti} могут либо задаваться экспериментатором («планируемый эксперимент»), или же быть случайными величинами («пассивный эксперимент»), но в любом случае предполагается, что погрешности ξ_t от них не зависят. Задача состоит в том, чтобы по известным значениям x_{ti} и y_t оценить параметры регрессии $\theta_0, \ldots, \theta_n$ и дисперсию погрешностей $\sigma^2 = \mathsf{D}\xi_t$. Часто для упрощения задачи дополнительно предполагают, что погрешности ξ_t нормально распределены.

Запишем модель (33.1) в матричной форме. Для этого определим векторы

$$\theta = (\theta_0, \theta_1, \ldots, \theta_n) \in \mathbb{R}^{n+1},$$

$$X_0 = (1, \ldots, 1) \in \mathbb{R}^N, \qquad \Xi = (\xi_1, \ldots, \xi_N) \in \mathbb{R}^N.$$

Из вектор-столбцов $X_0, X_1, \ldots, X_n$ составим матрицу X размерности $N \times (n+1)$. В этих обозначениях модель регрессии (33.1) принимает форму

$$Y = X\theta + \Xi. \tag{33.2}$$

Пусть $L(X)$ — линейная оболочка векторов $X_0, X_1, \ldots, X_n$:

$$L(X) = \ell(X_0, X_1, \ldots, X_n) = \bigl\{ X\theta \bigm| \theta \in \mathbb{R}^{n+1} \bigr\},$$

а P_X — ортогональный проектор пространства $\mathbb{R}^N$ на $L(X)$. Оценку $\hat\theta$ для вектора параметров θ в модели (33.2) ищут методом наименьших квадратов. Мы ее вычислили в предыдущем параграфе:

$$\hat{\theta} = (X^*X)^{-1}X^*Y. \tag{33.3}$$

Там же было установлено соотношение

$$P_XY = X\hat{\theta}.$$

Подставим выражение (33.2) в формулу (33.3):

$$\hat{\theta} = (X^*X)^{-1}X^*(X\theta + \Xi) = \theta + (X^*X)^{-1}X^*\Xi. \tag{33.4}$$

Поскольку мы предполагаем, что вектор погрешностей Ξ не зависит от X и имеет нулевое математическое ожидание, отсюда следует, что $\mathsf{E}\hat{\theta} = \theta$, то есть оценка $\hat{\theta}$ несмещенная. Заметим, что $\mathrm{Cov}\{\Xi, \Xi\} = \sigma^2 I$, где $\sigma^2 = \mathsf{D}\xi_t$. Поэтому в случае планируемого эксперимента (то есть когда матрица X фиксирована)

$$\begin{aligned}\mathrm{Cov}\{\hat{\theta}, \hat{\theta}\} &= \mathsf{E}\big\{\big(\hat{\theta} - \theta\big)\big(\hat{\theta} - \theta\big)^*\big\} = \mathsf{E}\big\{(X^*X)^{-1}X^*\Xi\Xi^*X(X^*X)^{-1}\big\} = \\ &= (X^*X)^{-1}X^*\big(\sigma^2 I\big)X(X^*X)^{-1} = \sigma^2(X^*X)^{-1}.\end{aligned} \tag{33.5}$$

Если еще погрешности ξ_t нормально распределены, то из (33.4) и (33.5) следует, что оценка $\hat{\theta}$ имеет распределение $\mathcal{N}_{n+1}\big(\theta, \sigma^2(X^*X)^{-1}\big)$.

Напомним, что *аффинным* подпространством называют сдвинутое линейное подпространство (сумму вида $v + L$, где v — фиксированный вектор, а L — какое-то линейное подпространство, проходящее через начало отсчета). Аффинные подпространства чаще всего задаются с помощью систем неоднородных линейных уравнений.

Рассмотрим модель регрессии (33.2) с фиксированной матрицей X и неизвестным вектором параметров θ. Очевидно, при нормально распределенных погрешностях ξ_t случайный вектор $\Xi = (\xi_1, \ldots, \xi_N)$ имеет распределение $\mathcal{N}_N(0, \sigma^2 I)$. Пусть в $L(X)$ задано аффинное подпространство l размерности m, а P_l — ортогональный проектор из $\mathbb{R}^N$ в l. Для проверки различных гипотез относительно параметров линейной регрессии полезна следующая теорема.

Теорема 33.1. *Если в модели линейной регрессии* (33.2) *вектор погрешностей Ξ имеет распределение $\mathcal{N}_N(0, \sigma^2 I)$ и все столбцы матрицы X линейно независимы, то*

а) случайная величина $|Y - P_XY|^2$ не зависит от $|P_XY - P_lY|^2$;

б) случайная величина $|Y - P_XY|^2$ распределена как $\sigma^2\chi^2_{N-n-1}$;

в) если $X\theta \in l$, то случайная величина $|P_XY - P_lY|^2$ распределена как $\sigma^2\chi^2_{n+1-m}$;

г) если $X\theta \in l$, то отношение

$$\frac{N-n-1}{n+1-m}\,\frac{|P_XY - P_lY|^2}{|Y - P_XY|^2} \tag{33.6}$$

имеет распределение Фишера $F_{n+1-m,N-n-1}$;

д) в частности, если $X\theta \in l$ и $m = n$, то величина

$$\sqrt{N-n-1}\,\frac{|P_XY - P_lY|}{|Y - P_XY|}$$

распределена как модуль случайной величины Стьюдента $|t_{N-n-1}|$.

Пусть $\hat{R}_{yx}$ — выборочная множественная корреляция между Y и совокупностью выборок $X_1, \ldots, X_n$.

Следствие 33.2. *Если в условиях предыдущей теоремы параметры $\theta_1, \ldots, \theta_n$ нулевые, то величина*

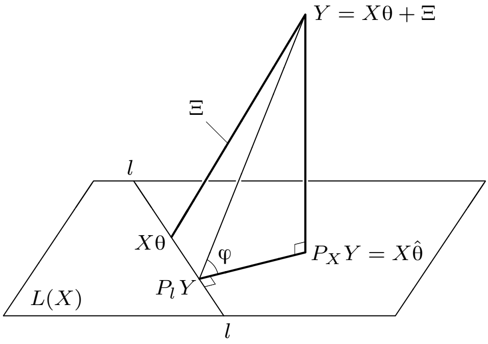
$$\frac{N-n-1}{n}\,\frac{\hat{R}^2_{yx}}{1-\hat{R}^2_{yx}} \tag{33.7}$$

имеет распределение Фишера $F_{n,N-n-1}$.

Доказательство теоремы 33.1 и следствия из нее изображено на рис. 1 и 2. По большому счету, все их утверждения вытекают из леммы о проекциях (леммы 25.1).

Произведем сдвиг пространства $\mathbb{R}^N$ на вектор $X\theta$. После него случайный вектор Y станет равен Ξ и будет иметь нормальное распределение $\mathcal{N}_N(0, \sigma^2 I)$, подпространство $L(X)$ перейдет в себя, подпространство l станет линейным (проходящим через начало отсчета), а рассматриваемые в теореме разности $Y - P_XY$ и $P_XY - P_lY$ вообще не изменятся. Поэтому достаточно доказать теорему для сдвинутого пространства.

По построению вектор P_XY является ортогональной проекцией Y на подпространство $L(X)$, вектор $Y - P_XY$ является проекцией Y на ортогональное дополнение к $L(X)$ в $\mathbb{R}^N$ (которое имеет размерность $N - n - 1$), а вектор $P_XY - P_lY$ является проекцией Y на ортогональное дополнение к подпространству l в пространстве $L(X)$ (имеющее



$$\dim L(X) = n + 1$$
$$\dim l = m$$
$$\frac{|P_X Y - P_l Y|^2}{|Y - P_X Y|^2} = \frac{\chi^2_{n+1-m}}{\chi^2_{N-n-1}}$$
если $m = n$, то
$$\frac{|P_X Y - P_l Y|}{|Y - P_X Y|} = \frac{|t_{N-n-1}|}{\sqrt{N-n-1}}$$

Рис. 1. Иллюстрация к теореме 33.1

размерность $n + 1 - m$). В этих условиях утверждения а), б), в) и г) вытекают из леммы о проекциях 25.1. Заметим, что случайные величины Фишера и Стьюдента связаны соотношением $F_{1,n} = t_n^2$. Поэтому при $m = n$ из г) вытекает д).

Чтобы доказать следствие 33.2, напомним формулу (31.5) для выборочной множественной корреляции:

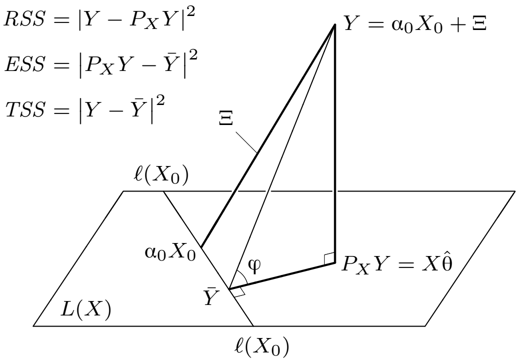
$$\hat{R}_{yx} = \frac{\left|P_X Y - \bar{Y}\right|}{\left|Y - \bar{Y}\right|}.$$

Из нее следует, что

$$\frac{\hat{R}^2_{yx}}{1 - \hat{R}^2_{yx}} = \frac{\left|P_X Y - \bar{Y}\right|^2}{\left|Y - \bar{Y}\right|^2 - \left|P_X Y - \bar{Y}\right|^2} = \frac{\left|P_X Y - \bar{Y}\right|^2}{\left|Y - P_X Y\right|^2}. \tag{33.8}$$

Пусть l — прямая, проходящая через вектор $X_0 = (1, \ldots, 1)$. Заметим, что $\bar{Y}$ — это проекция Y на прямую l, а условие $\theta_1 = \ldots = \theta_n = 0$ равносильно тому, что $X\theta \in l$. Сравнивая выражение (33.6) с (33.7) и (33.8), убеждаемся, что статистика (33.7) является частным случаем (33.6) при $m = 1$. $\square$

Приведем примеры проверки гипотез о параметрах множественной линейной регрессии. Пусть гипотеза H_0 состоит в том, что все параметры $\theta_1, \ldots, \theta_n$ нулевые, а гипотеза H_1 — в том, что хотя бы один из них отличен от нуля. Критерий для проверки этих гипотез называется



$$TSS = ESS + RSS$$

$$\cos\varphi = \hat{R}_{yx} = \frac{|P_X Y - \bar{Y}|}{|Y - \bar{Y}|}$$

$$(\operatorname{ctg}\varphi)^2 = \frac{|P_X Y - \bar{Y}|^2}{|Y - P_X Y|^2}$$

$$\frac{\hat{R}^2_{yx}}{1 - \hat{R}^2_{yx}} = \frac{\chi^2_n}{\chi^2_{N-n-1}}$$

Рис. 2. Иллюстрация к следствию 33.2

критерием значимости регрессии переменной y на совокупность переменных $x = (x_1, \ldots, x_n)$, или критерием некоррелированности y и x, или критерием независимости y от x (хотя последнее название корректно лишь в том случае, когда разность $y - P_x y$ нормально распределена и не зависит от x). Следствие 33.2 позволяет сформулировать его следующим образом:

$$\begin{cases} H_0, & \text{если } \dfrac{N-n-1}{n}\,\dfrac{\hat{R}^2_{yx}}{1-\hat{R}^2_{yx}} \leqslant \Delta, \\[2ex] H_1, & \text{если } \dfrac{N-n-1}{n}\,\dfrac{\hat{R}^2_{yx}}{1-\hat{R}^2_{yx}} > \Delta, \end{cases} \qquad \Delta = F^{-1}_{F_{n,N-n-1}}(1-\alpha).$$

Здесь Δ — квантиль уровня $1-\alpha$ для функции распределения Фишера $F_{F_{n,N-n-1}}$, определяемый по уровню значимости α. В частности, при $n = 1$ последний критерий можно записать в терминах распределения Стьюдента:

$$\begin{cases} H_0, & \text{если } \sqrt{N-2}\,\dfrac{\hat{R}_{yx}}{\left(1-\hat{R}^2_{yx}\right)^{1/2}} \leqslant \Delta, \\[2ex] H_1, & \text{если } \sqrt{N-2}\,\dfrac{\hat{R}_{yx}}{\left(1-\hat{R}^2_{yx}\right)^{1/2}} > \Delta, \end{cases} \qquad \Delta = F^{-1}_{t_{N-2}}\Bigl(1-\frac{\alpha}{2}\Bigr).$$

Пусть теперь гипотеза H_0 состоит в том, что параметр θ_n равен некоторому числу b, а гипотеза H_1 — в том, что $\theta_n \neq b$. Обозначим через l_0 линейную оболочку векторов $X_0, \ldots, X_{n-1}$, через l аффинное подпространство $bX_n + l_0$, а через P_l и P_{l_0} ортогональные проекторы на l и на l_0 соответственно. В этих обозначениях условие $\theta_n = b$ равносильно условию $X\theta \in l$. Если последнее условие выполняется, то в силу пункта д) теоремы 33.1 величина

$$\sqrt{N-n-1}\,\frac{|P_X Y - P_l Y|}{|Y - P_X Y|}$$

распределена как $|t_{N-n-1}|$. Нетрудно видеть, что

$$|P_X Y - P_l Y| = |\hat{\theta}_n - b|\,|X_n - P_{l_0} X_n|\,.$$

Для краткости записи положим $c = |X_n - P_{l_0} X_n|$ (это расстояние от вектора X_n до подпространства l_0). Тогда искомый критерий можно записать так:

$$\begin{cases} \theta_n = b, & \text{если} \quad \sqrt{N-n-1}\,\dfrac{|\hat{\theta}_n - b|\,c}{|Y - P_X Y|} \leqslant \Delta, \\[2ex] \theta_n \neq b, & \text{если} \quad \sqrt{N-n-1}\,\dfrac{|\hat{\theta}_n - b|\,c}{|Y - P_X Y|} > \Delta, \end{cases}$$

$$\Delta = F^{-1}_{t_{N-n-1}}\Big(1 - \frac{\alpha}{2}\Big).$$

Задача 1. Постройте доверительный интервал для θ_n.

Задача 2*. Докажите, что приведенные критерии состоятельны.

В силу пункта б) теоремы 33.1 случайная величина $|Y - P_X Y|^2$ распределена как $\sigma^2\chi^2_{N-n-1}$. Это позволяет взять в качестве несмещенной оценки для σ^2 остаточный средний квадрат

$$\hat{\sigma}^2 = \frac{|Y - P_X Y|^2}{N-n-1} = \frac{RSS}{N-n-1}.$$

Задача 3. Постройте доверительный интервал для σ^2 и критерий для проверки гипотезы $\sigma^2 = \sigma_0^2$.

Неожиданным приятным фактом является то, что в случае пассивного эксперимента теорема 33.1, следствие из нее и базирующиеся на них критерии остаются в силе. Основное отличие пассивного эксперимента от планируемого состоит в том, что в нем матрица X и линейное подпространство $L(X)$ становятся *случайными*. Поэтому нам следует более четко оговорить, каким образом определяется аффинное подпространство l в $L(X)$.

Зафиксируем в $\mathbb{R}^{n+1}$ какое-нибудь аффинное подпространство l_θ размерности m. Сопоставим ему аффинное подпространство l в $L(X)$ следующим образом:

$$l = \{X\theta \mid \theta \in l_\theta\}.$$

Если столбцы матрицы X линейно независимы, то l тоже имеет размерность m, а условие $X\theta \in l$ равносильно условию $\theta \in l_\theta$.

Теорема 33.3. *Если в модели линейной регрессии* (33.2) *вектор погрешностей* Ξ *не зависит от случайной матрицы* X, *имеет распределение* $\mathcal{N}_N(0, \sigma^2 I)$, *и к тому же* $\operatorname{rank} X = n + 1$ *почти наверное, то все утверждения теоремы 33.1 и следствия из нее остаются в силе* (*с заменой условия* $X\theta \in l$ *на* $\theta \in l_\theta$).

Доказательство. При каждом фиксированном значении матрицы X теорема 33.1 выполняется для *условного* распределения пары случайных величин $|Y - P_X Y|^2$ и $|P_X Y - P_l Y|^2$. А поскольку это условное распределение не зависит от X, оно совпадает с безусловным распределением указанной пары случайных величин (см. теорему 10.2 из первой части курса). $\square$

Глава 5. НЕКОТОРЫЕ ЗАДАЧИ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА

§ 34. Метод главных компонент

Для каждого случайного вектора $x = (x_1, \ldots, x_n)$ определяют математическое ожидание $\mu = \mathsf{E}x$, полную дисперсию $\mathsf{D}x = \mathsf{E}\,|x - \mu|^2$ и матрицу ковариаций $\Sigma = \mathsf{E}\{(x - \mu)(x - \mu)^*\}$. Очевидно,

$$\mathsf{D}x = \mathsf{E}\big\{(x_1 - \mu_1)^2 + \ldots + (x_n - \mu_n)^2\big\} = \mathsf{D}x_1 + \ldots + \mathsf{D}x_n.$$

Другими словами, дисперсия случайного вектора x равняется сумме дисперсий его декартовых координат.

Вариацией случайного вектора x относительно точки $a \in \mathbb{R}^n$ называется величина

$$\mathrm{Var}_a\, x = \mathsf{E}\,|x - a|^2\,.$$

Вариация и дисперсия связаны между собой тождеством

$$\mathrm{Var}_a\, x = \mathsf{D}x + |\mu - a|^2\,. \tag{34.1}$$

Действительно,

$$\mathsf{E}\,|x - a|^2 = \mathsf{E}\big\{\big((x - \mu) + (\mu - a), (x - \mu) + (\mu - a)\big)\big\} =$$
$$= \mathsf{E}\,|x - \mu|^2 + \mathsf{E}\,|\mu - a|^2 + 2\mathsf{E}\big\{(x - \mu, \mu - a)\big\} = \mathsf{D}x + |\mu - a|^2\,.$$

Из тождества (34.1) вытекает, что вариация становится минимальной при $a = \mu$; в этом случае она совпадает с полной дисперсией.

Ковариационная матрица Σ всегда симметрична и неотрицательно определена. Поэтому для нее существует ортонормированный базис, состоящий из собственных векторов $e_1, \ldots, e_n$:

$$\Sigma e_i = \lambda_i e_i, \qquad \lambda_i \geqslant 0.$$

Упорядочим этот базис по убыванию собственных значений λ_i, чтобы выполнялись неравенства $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$. Прямые, проходящие через векторы e_i, называют *главными осями,* а проекции случайного вектора x на главные оси — его *главными компонентами.* Очевидно, всякий вектор x раскладывается в сумму главных компонент:

$$x = (x, e_1)e_1 + \ldots + (x, e_n)e_n. \tag{34.2}$$

Довольно часто главными компонентами называют не проекции x на главные оси, а координаты (x, e_i), хотя последние определены лишь с точностью до знака.

Полная дисперсия случайного вектора равна сумме дисперсий его главных компонент:

$$\mathsf{E}\,|x-\mu|^2 = \mathsf{E}\left\{\sum_{i=1}^{n}(x-\mu, e_i)^2\right\} = \sum_{i=1}^{n}\mathsf{E}\big((x,e_i)-(\mu,e_i)\big)^2 = \sum_{i=1}^{n}\mathsf{D}(x, e_i).$$

Предложение 34.1. *Главные компоненты некоррелированы, а их дисперсии равны собственным значениям матрицы ковариаций.*

Доказательство. Не ограничивая общности, можно считать, что $\mu = 0$. Тогда

$$\mathrm{Cov}\{(x,e_i),(x,e_j)\} = \mathsf{E}\{e_i^* x \cdot x^* e_j\} = e_i^*\Sigma e_j = e_i^*\lambda_j e_j = \lambda_j\delta_{ij}. \quad \square$$

Главные компоненты используют для решения следующей задачи. Предположим, что мы хотим уменьшить размерность случайного вектора x, сохранив при этом максимальное количество информации о нем. Уменьшение размерности будем производить с помощью проектирования x на различные подпространства $L \subset \mathbb{R}^n$. Мерой информации, содержащейся в x и в проекции $P_L x$, будем считать дисперсии $\mathsf{D}x$ и $\mathsf{D}\{P_L x\}$ соответственно. Очевидно, максимально информативной одномерной проекцией будет первая главная компонента, максимально информативной двумерной проекцией будет сумма первых двух главных компонент и т. д. Если требуется найти проекцию минимальной размерности, содержащую долю не меньше $1-\varepsilon$ от всей информации, содержащейся в x, то для этого находят минимальное число k, при котором выполняется условие $\lambda_1 + \ldots + \lambda_k \geqslant (1-\varepsilon)\mathsf{D}x$, и в качестве искомой проекции берут сумму первых k главных компонент.

Довольно часто координаты случайного вектора x имеют разную физическую природу (длина, вес, время и т. п.), и главные компоненты существенно зависят от выбора соответствующих единиц измерения.

Поэтому они не имеют физического смысла и неадекватно отражают информативность разных признаков. В этой ситуации рекомендуется вначале *нормализовать* каждую координату x (сделать сдвиг и растяжение, приводящие к нулевому среднему и единичной дисперсии).

Все вышесказанное практически дословно переносится на случай многомерных выборок. Для этого достаточно отождествить каждую выборку $X = (x_1, \ldots, x_N)$, где $x_t \in \mathbb{R}^n$, с дискретной случайной величиной, принимающей значения $x_1, \ldots, x_N$ с равными вероятностями. Появляются только небольшие отличия в терминологии: математическое ожидание превращается в выборочное среднее $\bar{x}$, которое также называют *центром* выборки, матрица Σ превращается в выборочную матрицу ковариаций $\hat{\Sigma}$, вместо выборочной дисперсии $\mathsf{D}\{X\}$ рассматривают *полную сумму квадратов*

$$SS = N\,\mathsf{D}\{X\} = \sum_{t=1}^{N} |x_t - \bar{x}|^2 = \sum_{t=1}^{N}\sum_{i=1}^{n}(x_{ti} - \bar{x}_i)^2,$$

а вместо выборочных дисперсий главных компонент (x_t, e_i) начинают рассматривать частичные суммы квадратов $\sum_{t=1}^{N}(x_t - \bar{x}, e_i)^2$.

Задача. Докажите, что если L — линейное подпространство в $\mathbb{R}^n$ размерности k, то тогда $\mathsf{D}\{P_L x\} \leqslant \lambda_1 + \ldots + \lambda_k$, где λ_i — собственные значения матрицы ковариаций, упорядоченные по убыванию.

§ 35. Дискриминантный и кластерный анализы

Пусть на пространстве $\mathbb{R}^n$ есть несколько плотностей вероятностей $p_1(x), \ldots, p_k(x)$, которым отвечают априорные вероятности $\pi_1, \ldots, \pi_k$ (дающие в сумме единицу). Рассмотрим *плотность-смесь*

$$p(x) = \pi_1 p_1(x) + \ldots + \pi_k p_k(x).$$

Ее интерпретируют следующим образом: случайный вектор x с вероятностями π_i попадает в один из классов с номерами $i = 1, \ldots, k$, а внутри каждого класса распределение имеет плотность $p_i(x)$.

Предположим, что нам по значению x нужно определить, к какому классу оно относится. Конечно, если носители плотностей $p_i(x)$ пересекаются (а так чаще всего и бывает), то точно узнать номер класса невозможно, так как всегда будет ненулевая вероятность ошибиться. В этой ситуации целесообразно делать выбор таким образом, чтобы вероятность ошибки была минимальной. Соответствующее решающее правило называется *байесовским.*

В первой части данного курса мы доказали (см. пример 1 из § 20), что решающее правило является байесовским тогда и только тогда, когда оно сопоставляет каждому значению x тот номер класса i, при котором достигается максимум выражения

$$P_i(x) = \frac{\pi_i p_i(x)}{p(x)}.$$

Числа $P_i(x)$ называются апостериорными вероятностями классов (при заданном x). Значит, байесовское решающее правило рекомендует выбирать класс с наибольшей апостериорной вероятностью.

1. Дискриминантный анализ. Предположим, что априорные вероятности π_i и плотности $p_i(x)$ нам неизвестны, но имеется *обучающая выборка* $X = (x_1, \ldots, x_N)$, в которой для каждого значения x_t задан номер класса $i = d(x_t)$. По этой выборке требуется построить оценки для π_i, $p_i(x)$ и соответствующее байесовское решающее правило.

Обычно делают допущение, что внутри каждого класса вектор x имеет нормальное распределение $\mathcal{N}_n(\mu_i, \Sigma_i)$. Тогда в качестве оценок априорных вероятностей π_i берут относительное число выборочных значений, попавших в i-й класс:

$$\hat{\pi}_i = \frac{N_i}{N}, \qquad N_i = \#\{t \mid d(x_t) = i\}. \tag{35.1}$$

Оценками для μ_i и Σ_i служат выборочное среднее и несмещенная выборочная матрица ковариаций для каждого класса:

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{d(x_t)=i} x_t, \qquad S_i = \frac{1}{N_i - 1} \sum_{d(x_t)=i} (x_t - \hat{\mu}_i)(x_t - \hat{\mu}_i)^*. \tag{35.2}$$

Довольно часто рассматривают *модель Фишера,* в которой дополнительно предполагают, что все классы имеют одну и ту же матрицу ковариаций $\Sigma_i = \Sigma$. Тогда в качестве оценки для нее берут величину

$$S = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{d(x_t)=i} (x_t - \hat{\mu}_i)(x_t - \hat{\mu}_i)^* = \sum_{i=1}^{k} \frac{N_i - 1}{N - k} S_i. \tag{35.3}$$

Из несмещенности оценок S_i вытекает, что в случае $\Sigma_i = \Sigma$ оценка S для матрицы Σ тоже несмещенная.

Напомним, что плотность распределения $\mathcal{N}_n(\hat{\mu}_i, S_i)$ имеет вид

$$\hat{p}_i(x) = \frac{1}{(2\pi)^{n/2}\sqrt{|S_i|}} \exp\Big\{-\frac{1}{2}(x-\hat{\mu}_i)^* S_i^{-1}(x-\hat{\mu}_i)\Big\}. \tag{35.4}$$

По определению, подстановочное байесовское решающее правило относит вектор x в тот класс i, на котором достигается максимум выборочных апостериорных вероятностей

$$\hat{P}_i(x) = \frac{\hat{\pi}_i \hat{p}_i(x)}{\hat{p}(x)}, \qquad \hat{p}(x) = \sum_{i=1}^{k} \hat{\pi}_i \hat{p}_i(x). \tag{35.5}$$

Используя (35.4) и (35.5), его можно записать в виде

$$d(x) = \arg\min_{1\leqslant i\leqslant k} \big\{(x-\hat{\mu}_i)^* S_i^{-1}(x-\hat{\mu}_i) + \ln|S_i| - 2\ln\hat{\pi}_i\big\}. \tag{35.6}$$

В частности, в модели Фишера (когда все S_i заменяются на S) это решающее правило принимает форму

$$d(x) = \arg\min_{1\leqslant i\leqslant k} \big\{\hat{\mu}_i^* S^{-1}\hat{\mu}_i - 2\hat{\mu}_i^* S^{-1}x - 2\ln\hat{\pi}_i\big\}. \tag{35.7}$$

Дискриминантный анализ на основе формулы (35.6) иногда называют *квадратичным*, а на основе формулы (35.7) — *линейным* (потому что в них участвуют соответственно квадратичные и линейные по x функции).

2. Кластерный анализ. Задача кластерного анализа состоит в том, чтобы разбить множество значений выборки $X = (x_1, \ldots, x_N)$ на несколько классов так, чтобы элементы внутри каждого класса были бы в каком-то смысле близки, а элементы из разных классов имели бы больше различий между собой. При этом возможны два типа классификации: простая, когда разные классы не пересекаются, и иерархическая, когда некоторые классы могут быть вложены друг в друга (например, классификация живых существ в биологии).

Для построения иерархической классификации (часто называемой *таксономией*) следует выбрать какой-то способ измерения расстояний между выборочными значениями, а также между разными классами.

Если каждый элемент выборки принадлежит $\mathbb{R}^n$ (то есть описывается набором n признаков), то в качестве расстояния между выборочными значениями x и y можно взять, например, обычное евклидово расстояние $\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$, или манхэттенское расстояние $\sum_{i=1}^n |x_i - y_i|$, или расстояние Чебышёва $\max_i |x_i - y_i|$, или количество всех несовпавших признаков $\#\{i \mid x_i \neq y_i\}$ (последнее используют, когда признаки принимают лишь конечное число значений). Расстояние между классами можно определить как расстояние между наиболее близкими точками из этих классов, как расстояние между самыми далекими точками из этих классов, как расстояние между центрами классов, как различные взвешенные средние от расстояний между точками классов и т. п. Обычно координаты выборочных значений (признаки) имеют различную физическую природу или разный масштаб измерений. Поэтому их рекомендуется предварительно нормализовать (чтобы выборочное среднее по каждой координате сделалось нулевым, а выборочная дисперсия единичной).

Затем осуществляется пошаговая кластеризация. На нулевом шаге каждое выборочное значение образует свой индивидуальный класс. На каждом следующем шаге находятся два ближайших класса и объединяются в один класс следующего уровня. Этот процесс «склеивания» классов происходит до тех пор, пока они в конце концов не образуют один максимальный класс, содержащий всю выборку.

В итоге будет построено иерархическое дерево классов. Потом это дерево можно агрегировать: разбить весь диапазон расстояний между классами на несколько интервалов, и в пределах каждого интервала оставить только максимальные по включению классы. Этим классам стараются приписать какой-нибудь содержательный смысл и дать подходящие названия.

Иначе действуют, когда требуется разбить выборку на небольшое число непересекающихся классов. В этом случае применяют итерационную процедуру, которая базируется на дискриминантном анализе и называется методом k-средних. Опишем ее.

Предположим, что нужно разбить выборку $X = (x_1, \ldots, x_N)$ на k классов. На нулевом шаге следует выбрать какие-нибудь k точек $\hat{\mu}_i^{(0)}$, $i = 1, \ldots, k$, в качестве начальных приближений к центрам классов. Начальным приближением для ковариационной матрицы объявляют единичную матрицу $S^{(0)} = I$. На каждом следующем шаге все выборочные значения классифицируют по степени их близости (в метрике

Махаланобиса) к центрам классов, полученным на предыдущем шаге:

$$d^{(l)}(x_t) = \arg\min_{1\leqslant i\leqslant k}\Big(x_t - \hat{\mu}_i^{(l-1)}\Big)^*\Big(S^{(l-1)}\Big)^{-1}\Big(x_t - \hat{\mu}_i^{(l-1)}\Big), \qquad (35.8)$$

а также пересчитывают центры классов

$$\hat{\mu}_i^{(l)} = \frac{1}{N_i^{(l)}}\sum_{d^{(l)}(x_t)=i} x_t, \quad \text{где} \quad N_i^{(l)} = \#\big\{t \mid d^{(l)}(x_t) = i\big\},$$

и выборочную ковариационную матрицу (в модели Фишера)

$$S^{(l)} = \frac{1}{N-k}\sum_{i=1}^{k}\left\{\sum_{d^{(l)}(x_t)=i}\Big(x_t - \hat{\mu}_i^{(l)}\Big)\Big(x_t - \hat{\mu}_i^{(l)}\Big)^*\right\}.$$

Процедура заканчивается в тот момент, когда классы перестают изменяться.

Эффективность описанного метода k-средних существенно зависит от того, насколько правильно выбрано число классов и их начальные центры. Вместо решающего правила (35.8) можно использовать байесовское решающее правило (35.7) или (35.6), но это в меньшей степени влияет на качество кластерного анализа.

§ 36. Статистика T^2 Хотеллинга

Эту статистику используют для проверки гипотез о значении математического ожидания многомерного нормального распределения при неизвестной матрице ковариаций. Она служит обобщением статистики Стьюдента (точнее, ее квадрата) на многомерный случай.

Пусть имеется выборка $X = (x_1, \ldots, x_N)$, в которой $x_i \in \mathbb{R}^n$. Тогда *статистикой T^2 Хотеллинга* называют функцию

$$T^2 = T^2(X, \mu) = N(\bar{x} - \mu)^* S^{-1}(\bar{x} - \mu), \qquad (36.1)$$

где S — несмещенная выборочная матрица ковариаций:

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^*.$$

В одномерном случае (то есть при $n = 1$) эта статистика совпадает с квадратом статистики Стьюдента

$$t = \sqrt{N}\,\frac{\bar{x} - \mu}{s}, \qquad \text{где } s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

Теорема 36.1. *Если X — выборка мощности $N > n$ из невырожденного нормального распределения $\mathcal{N}_n(\mu, \Sigma)$, то статистика*

$$F_{n,N-n} = \frac{N-n}{n(N-1)}\,T^2(X, \mu)$$

есть случайная величина Фишера с n и $N - n$ степенями свободы.

Доказательство. В силу теоремы 26.1 вектор $v = \sqrt{N}(\bar{x} - \mu)$ имеет нормальное распределение $\mathcal{N}_n(0, \Sigma)$, а матрица $A = (N-1)S$ не зависит от v и имеет распределение Уишарта $\mathcal{W}_n(\Sigma, N-1)$. Поэтому утверждение теоремы вытекает из следующей леммы.

Лемма 36.2. *Если случайный вектор v имеет нормальное распределение $\mathcal{N}_n(0, \Sigma)$, где $|\Sigma| \neq 0$, а случайная матрица A не зависит от v и имеет распределение Уишарта $\mathcal{W}_n(\Sigma, N-1)$, где $N > n$, то тогда*

$$\frac{N-n}{n}\,v^* A^{-1} v = F_{n,N-n}.$$

Доказательство. В условиях леммы существуют такие независимые случайные векторы $z_1, \ldots, z_N$ с распределением $\mathcal{N}_n(0, \Sigma)$, что

$$z_1 = v, \qquad A = \sum_{t=2}^{N} z_t z_t^*.$$

Положим $y_t = \Sigma^{-1/2} z_t$. Очевидно, случайные векторы y_t независимы. Они имеют распределение $\mathcal{N}_n(0, I)$, поскольку $\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I$. Определим матрицу Уишарта

$$W = \sum_{t=2}^{N} y_t y_t^* = \sum_{t=2}^{N} \Sigma^{-1/2} z_t z_t^* \Sigma^{-1/2} = \Sigma^{-1/2} A \Sigma^{-1/2}. \tag{36.2}$$

Для нее выполняются равенства

$$v^* A^{-1} v = z_1^* \Sigma^{-1/2}\big(\Sigma^{1/2} A^{-1} \Sigma^{1/2}\big)\Sigma^{-1/2} z_1 = y_1^* W^{-1} y_1,$$

которые можно записать в равносильной форме

$$v^*A^{-1}v = |y_1|^2 e^* W^{-1} e, \quad \text{где} \quad e = \frac{y_1}{|y_1|}. \tag{36.3}$$

По построению случайная величина $|y_1|^2 = y_{11}^2 + \ldots + y_{1n}^2$ имеет распределение χ_n^2. Теперь нам достаточно доказать, что условное распределение выражения $e^* W^{-1} e$ при любом фиксированном единичном векторе e совпадает с распределением $1/\chi_{N-n}^2$. В самом деле, если это так, то тогда по теореме 10.2 случайная величина $e^* W^{-1} e$ не зависит от y_1, откуда следует, что произведение $|y_1|^2 e^* W^{-1} e$ распределено как случайная величина χ_n^2/χ_{N-n}^2, и тем самым лемма 36.2 доказана.

Фиксируем любой единичный вектор e. Пусть C — ортогональный поворот в $\mathbb{R}^n$, который переводит e в базисный орт $e_1 = (1, 0, \ldots, 0)$. Рассмотрим случайные векторы $u_t = C y_t$. Очевидно, они независимы и имеют распределение $\mathcal{N}_n(0, CIC^*) = \mathcal{N}_n(0, I)$. Положим

$$V = \sum_{t=2}^{N} u_t u_t^* = \sum_{t=2}^{N} C y_t y_t^* C^* = CWC^*. \tag{36.4}$$

По построению случайная матрица V имеет распределение Уишарта $\mathcal{W}_n(I, N-1)$. Для нее

$$e^* W^{-1} e = e^* C^* \bigl(C W^{-1} C^*\bigr) C e = e_1^* V^{-1} e_1 = \bigl(V^{-1}\bigr)_{11}.$$

Образуем из строк $u_t = (u_{t1}, \ldots, u_{tn})$, где $t = 2, \ldots, N$, матрицу U размерности $(N-1) \times n$. Тогда ее столбцы $U_j = (u_{2j}, \ldots, u_{Nj})$ имеют нормальное распределение $\mathcal{N}_{N-1}(0, I)$ и независимы. В силу формулы (36.4) элементами матрицы V являются скалярные произведения

$$v_{ij} = \sum_{t=2}^{N} u_{ti} u_{tj} = (U_i, U_j).$$

В линейной алгебре матрицу такого вида называют *матрицей Грама* для набора векторов $U_1, \ldots, U_n$ и обозначают $G(U_1, \ldots, U_n)$. Главное свойство этой матрицы заключается в том, что ее определитель равен квадрату n-мерного объема параллелепипеда, натянутого на векторы $U_1, \ldots, U_n$ (см. приложение, п. II). Кроме того, как нетрудно заметить, алгебраическое дополнение к элементу v_{11} в матрице V равно определителю матрицы Грама $G(U_2, \ldots, U_n)$.

Найдем элемент обратной матрицы $\left(V^{-1}\right)_{11}$ по формуле Крамера:

$$\left(V^{-1}\right)_{11} = \frac{|G(U_2, \ldots, U_n)|}{|V|}.$$

Обозначим через PU_1 ортогональную проекцию вектора $U_1 \in \mathbb{R}^{N-1}$ на линейную оболочку $\ell(U_2, \ldots, U_n)$. Объем параллелепипеда, натянутого на совокупность векторов $U_1, \ldots, U_n$, равен произведению объема его основания (в качестве которого мы возьмем параллелепипед, натянутый на векторы $U_2, \ldots, U_n$) и высоты $|U_1 - PU_1|$. Поэтому

$$\frac{|V|}{|G(U_2, \ldots, U_n)|} = |U_1 - PU_1|^2 .$$

Заметим, что вектор $U_1 - PU_1$ является ортогональной проекцией U_1 на ортогональное дополнение к подпространству $\ell(U_2, \ldots, U_n)$ в $\mathbb{R}^{N-1}$. В силу леммы о проекциях 25.1 квадрат его длины имеет распределение случайной величины χ^2_{N-n}. Значит, выражение $e^* W^{-1} e = (V^{-1})_{11}$ распределено как $1/\chi^2_{N-n}$. $\square$

Зная распределение статистики Хотеллинга, несложно построить критерий для проверки гипотезы о значении вектора математического ожидания μ при неизвестной матрице ковариаций Σ. Он имеет вид

$$\begin{cases} \mu = \mu_0, & \text{если } \; T^2(X, \mu_0) \leqslant \Delta, \\ \mu \neq \mu_0, & \text{если } \; T^2(X, \mu_0) > \Delta, \end{cases} \qquad \Delta = \frac{n(N-1)}{N-n} F^{-1}_{F_{n,N-n}}(1-\alpha), \tag{36.5}$$

где α — требуемый уровень значимости.

Можно также построить *доверительную область* уровня $1 - \alpha$ для истинного значения μ. Она определяется неравенством

$$N(\bar{x} - \mu)^* S^{-1} (\bar{x} - \mu) \leqslant \Delta$$

и представляет собой эллипсоид с центром в точке $\bar{x}$, имеющий линейные размеры порядка $1\big/\sqrt{N}$.

Пример. Пусть $X = (x_1, \ldots, x_N)$ — выборка из нормального распределения $\mathcal{N}_n(\mu, \Sigma)$ с неизвестными параметрами μ и Σ. Построить критерий для проверки гипотезы $A\mu = b$, где $A \colon \mathbb{R}^n \to \mathbb{R}^m$ — линейное отображение ранга m и $b \in \mathbb{R}^m$.

Решение. Определим случайные векторы $y_t = Ax_t$. Они нормально распределены и имеют математическое ожидание $A\mu$. Поэтому критерий для проверки гипотезы $A\mu = b$ полностью аналогичен (36.5):

$$\begin{cases} A\mu = b, & \text{если}\ \ T^2(Y,b) \leqslant \Delta, \\ A\mu \neq b, & \text{если}\ \ T^2(Y,b) > \Delta, \end{cases} \qquad \Delta = \frac{m(N-1)}{N-m}\, F^{-1}_{F_{m,N-m}}(1-\alpha),$$

где $Y = (y_1, \ldots, y_N)$.

Теорема 36.3. *Если X — выборка из нормального распределения $\mathcal{N}_n(\mu,\Sigma)$ мощности $N > n$, то отношение правдоподобия для проверки гипотезы $\mu = \mu_0$ при неизвестной матрице Σ имеет вид*

$$\Lambda(X) = \frac{\sup\bigl\{L(\mu,\Sigma) \bigm| \mu \in \mathbb{R}^n,\ \ \Sigma = \Sigma^* > 0\bigr\}}{\sup\bigl\{L(\mu_0,\Sigma) \bigm| \Sigma = \Sigma^* > 0\bigr\}} = \left(1 + \frac{T^2(X,\mu_0)}{N-1}\right)^{N/2},$$

где $L(\mu,\Sigma)$ — функция правдоподобия выборки X.

Задача. Докажите эту теорему с помощью леммы 27.2 и тождества $|I + vv^*| = 1 + v^*v$, в котором I — единичная матрица, а v — произвольный вектор-столбец.

§ 37. Сравнение математических ожиданий

1. Выборки с одинаковыми матрицами ковариаций. Пусть в пространстве $\mathbb{R}^n$ есть две независимые выборки $X = (x_1, \ldots, x_N)$ и $Y = (y_1, \ldots, y_M)$ из нормальных распределений с математическими ожиданиями μ_x и μ_y соответственно и одинаковыми матрицами ковариаций Σ. Предполагается, что величины μ_x, μ_y и Σ нам неизвестны. Требуется проверить гипотезу $\mu_x = \mu_y$ против альтернативы общего вида $\mu_x \neq \mu_y$.

Сделаем это при помощи T^2-статистики Хотеллинга. Рассмотрим статистики

$$z = \sqrt{\frac{NM}{N+M}}\,(\bar{x} - \bar{y}),$$

$$A_{xx} = \sum_{t=1}^{N} (x_t - \bar{x})(x_t - \bar{x})^*, \qquad A_{yy} = \sum_{t=1}^{M} (y_t - \bar{y})(y_t - \bar{y})^*,$$

$$A = A_{xx} + A_{yy}.$$

По теореме 26.1 случайные величины $\bar{x}$ и A_{xx} независимы и имеют распределения $\mathcal{N}_n\bigl(\mu_x, \frac{1}{N}\Sigma\bigr)$ и $\mathcal{W}_n(\Sigma, N-1)$. Аналогично, случайные величины $\bar{y}$ и A_{yy} независимы и имеют распределения $\mathcal{N}_n\big(\mu_y, \frac{1}{M}\Sigma\big)$ и $\mathcal{W}_n(\Sigma, M-1)$. Значит, случайные величины z и A тоже независимы. Поскольку $\bar{x}$ и $\bar{y}$ независимы,

$$\mathrm{Cov}\{\bar{x}-\bar{y}, \bar{x}-\bar{y}\} = \mathrm{Cov}\{\bar{x}, \bar{x}\} + \mathrm{Cov}\{\bar{y}, \bar{y}\} = \frac{1}{N}\Sigma + \frac{1}{M}\Sigma,$$

откуда следует, что $\mathrm{Cov}\{z, z\} = \Sigma$.

В итоге получается, что случайный вектор z имеет распределение $\mathcal{N}_n(\mu_z, \Sigma)$, в то время как случайная матрица A по построению имеет распределение $\mathcal{W}_n(\Sigma, N+M-2)$. Если выполняется гипотеза $\mu_x = \mu_y$, то тогда $\mu_z = 0$. В этом случае по лемме 36.2

$$\frac{N+M-n-1}{n} z^* A^{-1} z = F_{n, N+M-n-1}.$$

Поэтому решающее правило с уровнем значимости α можно записать в виде

$$\begin{cases} \mu_x = \mu_y, & \text{если} \quad z^* A^{-1} z \leqslant \Delta, \\ \mu_x \neq \mu_y, & \text{если} \quad z^* A^{-1} z > \Delta, \end{cases} \qquad \Delta = n\,\frac{F^{-1}_{F_{n,N+M-n-1}}(1-\alpha)}{N+M-n-1}.$$

Без доказательства отметим еще, что оно равносильно критерию отношения правдоподобия для проверки сложной гипотезы $\mu_x = \mu_y$.

2. Две выборки одинакового объема. Пусть в пространстве $\mathbb{R}^n$ имеются две независимые выборки $X = (x_1, \ldots, x_N)$ и $Y = (y_1, \ldots, y_N)$ из распределений $\mathcal{N}_n(\mu_x, \Sigma_{xx})$ и $\mathcal{N}_n(\mu_y, \Sigma_{yy})$ с неизвестными ковариационными матрицами Σ_{xx} и Σ_{yy}. Нужно проверить гипотезу $\mu_x = \mu_y$.

Положим $z_t = x_t - y_t$ и построим новую выборку $Z = (z_1, \ldots, z_N)$. Все случайные величины z_t независимы и имеют нормальное распределение $\mathcal{N}_n(\mu_x - \mu_y, \Sigma_{xx} + \Sigma_{yy})$. Поэтому гипотезу $\mu_x - \mu_y = 0$ можно проверить с помощью критерия (36.5), который в данном случае принимает форму

$$\begin{cases} \mu_x = \mu_y, & \text{если} \quad T^2 \leqslant \Delta, \\ \mu_x \neq \mu_y, & \text{если} \quad T^2 > \Delta, \end{cases} \qquad \Delta = \frac{n(N-1)}{N-n} F^{-1}_{F_{n,N-n}}(1-\alpha), \tag{37.1}$$

где

$$T^2 = N \bar{z}^* S^{-1} \bar{z}, \qquad S = \frac{1}{N-1} \sum_{t=1}^{N} (z_t - \bar{z})(z_t - \bar{z})^*. \tag{37.2}$$

3. Две произвольные выборки. Пусть заданы две выборки $X = (x_1, \ldots, x_N)$ и $Y = (y_1, \ldots, y_M)$ соответственно из нормальных распределений $\mathcal{N}_n(\mu_x, \Sigma_{xx})$ и $\mathcal{N}_n(\mu_y, \Sigma_{yy})$. Будем считать, что $N < M$. В этом случае для проверки гипотезы $\mu_x = \mu_y$ рекомендуется построить новую выборку $Z = (z_1, \ldots, z_N)$:

$$z_t = x_t - \sqrt{\frac{N}{M}}\left(y_t - \frac{1}{N}\sum_{s=1}^{N} y_s\right) - \frac{1}{M}\sum_{s=1}^{M} y_s. \tag{37.3}$$

Можно доказать, что определенные таким образом случайные векторы z_t независимы и имеют распределение $\mathcal{N}_n\big(\mu_x - \mu_y,\, \Sigma_{xx} + \frac{N}{M}\Sigma_{yy}\big)$. Поэтому гипотезу $\mu_x = \mu_y$ можно проверять с помощью полученного в предыдущем пункте критерия (37.1), (37.2).

Задача 1. Докажите, что случайные векторы z_t из (37.3) независимы и имеют указанное распределение.

Подсказка: проверьте, что $\mathrm{Cov}\{z_t, z_s\} = \delta_{ts}\big(\Sigma_{xx} + \frac{N}{M}\Sigma_{yy}\big)$.

4. Дисперсионный анализ. Пусть имеется m выборок разной мощности $X_i = (x_{i1}, \ldots, x_{iN_i})$, где $i = 1, \ldots, m$. Предполагается, что они взяты из одномерных нормальных распределений $\mathcal{N}_1(\mu_i, \sigma^2)$ с неизвестными математическими ожиданиями μ_i и неизвестными, но одинаковыми дисперсиями σ^2. Требуется проверить гипотезу о том, что все математические ожидания μ_i совпадают.

Рассмотрим составную выборку

$$X = (x_{11}, \ldots, x_{1N_1}, \quad . \; . \; . \; . \quad, x_{m1}, \ldots, x_{mN_m}).$$

Она имеет мощность $N = N_1 + \ldots + N_m$. Введем обозначения

$$\bar{x}_i = \frac{1}{N_i}\sum_{t=1}^{N_i} x_{it}, \qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{m}\sum_{t=1}^{N_i} x_{it}.$$

Величины $\bar{x}_i$ называются внутригрупповыми средними, а $\bar{x}$ — общим выборочным средним. Определим векторы

$$e_1 = (1, \ldots, 1, 0, \quad . \; . \; . \; . \quad, 0), \quad \ldots, \quad e_m = (0, \quad . \; . \; . \; . \quad, 0, 1, \ldots, 1)$$

(у вектора e_i единицы стоят на тех же позициях, что и переменные $x_{i1}, \ldots, x_{iN_i}$ в составной выборке X), а также

$$e = (1, \ldots, 1, \quad . \; . \; . \; . \quad, 1, \ldots, 1) = e_1 + \ldots + e_m.$$

Обозначим через L линейную оболочку векторов $e_1, \ldots, e_m$, через l прямую, содержащую e, а через P_L и P_l — ортогональные проекторы пространства $\mathbb{R}^n$ на L и на l соответственно. Легко видеть, что

$$P_L X = \bar{x}_1 e_1 + \ldots + \bar{x}_m e_m, \qquad P_l X = \bar{x} e$$

(проверьте это самостоятельно).

Применим лемму о проекциях к случайному вектору X. По условию он имеет нормальное распределение с матрицей ковариаций $\sigma^2 I$. Рассмотрим проекции X на ортогональное дополнение к подпространству L в $\mathbb{R}^n$ и на ортогональное дополнение к прямой l в L. Их можно записать как $X - P_L X$ и $P_L X - P_l X$ соответственно. Очевидно,

$$\mathsf{E}\{X - P_L X\} = \sum_{i=1}^{m} \mu_i e_i - \sum_{i=1}^{m} \mu_i e_i = 0,$$

а если к тому же $\mu_1 = \ldots = \mu_m = \mu$, то

$$\mathsf{E}\{P_L X - P_l X\} = \sum_{i=1}^{m} \mu_i e_i - \mu e = 0.$$

По лемме о проекциях случайные векторы $X - P_L X$ и $P_L X - P_l X$ не зависят друг от друга, случайная величина $|X - P_L X|^2$ распределена как $\sigma^2 \chi^2_{N-m}$, а если еще все μ_i совпадают, то тогда случайная величина $|P_L X - P_l X|^2$ распределена как $\sigma^2 \chi^2_{m-1}$, и отношение

$$F_{m-1, N-m} = \frac{N-m}{m-1} \frac{|P_L X - P_l X|^2}{|X - P_L X|^2} \tag{37.4}$$

имеет распределение Фишера с $m-1$ и $N-m$ степенями свободы.

Полученные результаты позволяют сформулировать критерий для проверки гипотезы $H_0\colon \mu_1 = \ldots = \mu_m$ (с уровнем значимости α):

$$\begin{cases} H_0, & \text{если } \dfrac{N-m}{m-1} \dfrac{|P_L X - P_l X|^2}{|X - P_L X|^2} \leqslant \Delta, \\[2ex] H_1, & \text{если } \dfrac{N-m}{m-1} \dfrac{|P_L X - P_l X|^2}{|X - P_L X|^2} > \Delta, \end{cases} \qquad \Delta = F^{-1}_{F_{m-1, N-m}}(1-\alpha).$$

Вычислим $|X - P_L X|^2$ и $|P_L X - P_l X|^2$ в явном виде:

$$|X - P_L X|^2 = \sum_{i=1}^{m} \sum_{t=1}^{N_i} (x_{it} - \bar{x}_i)^2, \qquad |P_L X - P_l X|^2 = \sum_{i=1}^{m} N_i (\bar{x}_i - \bar{x})^2.$$

Величину $|X - P_L X|^2$ часто называют внутригрупповой суммой квадратов, величину $|P_L X - P_l X|^2$ называют межгрупповой суммой квадратов, а все выражение (37.4) понимают как отношение межгрупповой выборочной дисперсии к внутригрупповой выборочной дисперсии, что и породило название «дисперсионный анализ».

Задача 2. Проверьте, что если сложить вместе внутригрупповую и межгрупповую суммы квадратов, то в результате получится полная сумма квадратов $\sum_{i,t} (x_{it} - \bar{x})^2$.

Задача 3. Вычислите доверительные интервалы для математических ожиданий μ_i и для разностей $\mu_i - \mu_j$.

5. Многомерный дисперсионный анализ. Пусть, как и прежде, имеется m независимых выборок $X_i = (x_{i1}, \ldots, x_{iN_i})$, где $i = 1, \ldots, m$, только на этот раз мы будем предполагать, что они взяты из многомерных распределений $\mathcal{N}_n(\mu_i, \Sigma)$ с неизвестными математическими ожиданиями μ_i и общей матрицей ковариаций Σ. Требуется проверить гипотезу о совпадении всех μ_i.

Рассмотрим составную выборку мощности $N = N_1 + \ldots + N_m$:

$$X = (x_{11}, \ldots, x_{1N_1}, \quad . \quad . \quad . \quad , x_{m1}, \ldots, x_{mN_m}).$$

Вычислим для нее внутригрупповые и общее выборочные средние

$$\bar{x}_i = \frac{1}{N_i} \sum_{t=1}^{N_i} x_{it}, \qquad \bar{x} = \frac{1}{N} \sum_{i=1}^{m} \sum_{t=1}^{N_i} x_{it},$$

а также ненормированные матрицы ковариаций

$$A_i = \sum_{t=1}^{N_i} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)^*, \qquad A = \sum_{i=1}^{m} A_i,$$

$$B = \sum_{i=1}^{m} \sum_{t=1}^{N_i} (x_{it} - \bar{x})(x_{it} - \bar{x})^*$$

(как обычно, мы считаем x_{it} вектор-столбцами размерности n). По теореме 26.1 матрицы A_i имеют распределения $\mathcal{W}_n(\Sigma, N_i - 1)$. Значит, матрица A распределена по закону $\mathcal{W}_n(\Sigma, N - m)$. А в случае $\mu_1 = \ldots = \mu_m$ матрица B имеет распределение Уишарта $\mathcal{W}_n(\Sigma, N - 1)$.

Предложение 37.1. *Если $\mu_1 = \ldots = \mu_m$ и $|\Sigma| \neq 0$, то отношение $\lambda = |A| / |B|$ является статистикой Уилкса $\lambda_n(N - m, m - 1)$.*

Поэтому критерий для проверки гипотезы $H_0: \mu_1 = \ldots = \mu_m$ можно сформулировать так:

$$\begin{cases} H_0, & \text{если } |A|/|B| \leqslant \Delta, \\ H_1, & \text{если } |A|/|B| > \Delta, \end{cases} \qquad \Delta = F^{-1}_{\lambda_n(N-m,m-1)}(1-\alpha).$$

Здесь предполагается, что выполняется условие $N - m \geqslant n$ (в противном случае статистика Уилкса тождественно обращается в нуль, и ее не рассматривают).

Докажем предложение 37.1. Как и в предыдущем подразделе, определим в $\mathbb{R}^N$ векторы

$$e_1 = (1, \ldots, 1, 0, \quad . \quad . \quad . \quad , 0), \quad \ldots, \quad e_m = (0, \quad . \quad . \quad . \quad , 0, 1, \ldots, 1),$$

$$e = (1, \ldots, 1, \quad . \quad . \quad . \quad , 1, \ldots, 1) = e_1 + \ldots + e_m.$$

Обозначим буквой L линейную оболочку векторов $e_1, \ldots, e_m$, через l — прямую, содержащую вектор e, а через P_L и P_l — ортогональные проекторы на L и на l.

Составим из всех вектор-строк x^*_{it} (упорядоченных, как и в выборке X) одну матрицу X размерности $N \times n$. Тогда для нее выполняются равенства

$$P_l X = e\bar{x}^*, \qquad P_L X = e_1\bar{x}^*_1 + \ldots + e_m\bar{x}^*_m$$

(в которых $P_l X$, $e\bar{x}^*$, $P_L X$, $e_i\bar{x}^*_i$ — матрицы размерности $N \times n$), а также

$$B = X^*(I - P_l)X, \qquad A = X^*(I - P_L)X$$

(проверьте эти равенства самостоятельно). Очевидно,

$$\mathsf{E}\{(I - P_L)X\} = \sum_{i=1}^m e_i\mu^*_i - \sum_{i=1}^m e_i\mu^*_i = 0,$$

а если $\mu_1 = \ldots = \mu_m = \mu$, то

$$\mathsf{E}\{(P_L - P_l)X\} = \sum_{i=1}^m e_i\mu^* - e\mu^* = 0.$$

Применим лемму Андерсона 25.4 к матрице X и паре проекторов $P_1 = I - P_L$ и $P_2 = P_L - P_l$. По этой лемме отношение

$$\lambda = \frac{|X^* P_1 X|}{\bigl|X^*(P_1 + P_2)X\bigr|} = \frac{|A|}{|B|}$$

является статистикой Уилкса $\lambda_n(N - m, m - 1)$. $\square$

Задача 4. Докажите, что отношение правдоподобия для проверки гипотезы $H_0: \mu_1 = \ldots = \mu_m$ имеет вид $\Lambda = \lambda^{-N/2}$, где $\lambda = |A|/|B|$.

Задача 5. Докажите, что для статистики Уилкса $\lambda = \lambda_n(N-1, 1)$ выполняется тождество

$$\lambda^{-1} = 1 + \frac{\chi^2_n}{\chi^2_{N-n}}.$$

§ 38. Проверка гипотезы о независимости нескольких групп переменных†

Пусть случайный вектор $x \in \mathbb{R}^n$ имеет невырожденное нормальное распределение $\mathcal{N}_n(\mu, \Sigma)$, и совокупность всех его координат разбита на k блоков: $x = \bigl(x^{(1)}, \ldots, x^{(k)}\bigr)$, где $x^{(i)} \in \mathbb{R}^{n_i}$ и $n_1 + \ldots + n_k = n$. Соответственно математическое ожидание μ и матрица ковариаций Σ разбиваются на блоки

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \ldots & \Sigma_{1k} \\ \Sigma_{21} & \Sigma_{22} & \ldots & \Sigma_{2k} \\ \vdots & \vdots & & \vdots \\ \Sigma_{k1} & \Sigma_{k2} & \ldots & \Sigma_{kk} \end{pmatrix},$$

где $\mu_i = \mathsf{E} x^{(i)}$ и $\Sigma_{ij} = \mathrm{Cov}\bigl\{x^{(i)}, x^{(j)}\bigr\}$. По выборке $X = (x_1, \ldots, x_N)$ из этого распределения требуется проверить гипотезу о том, что блоки координат $x^{(i)}$ не зависят друг от друга.

Известно, что независимость случайных векторов $x^{(i)}$ равносильна их некоррелированности: $\Sigma_{ij} = 0$ при $i \neq j$. Обозначим гипотезу независимости через H_0, а ее альтернативу через H_1. Для проверки сложной гипотезы H_0 используем критерий отношения правдоподобия из § 19 первой части курса.

Функция правдоподобия для выборки X имеет вид

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{nN/2} \, |\Sigma|^{N/2}} \exp\biggl\{-\frac{1}{2} \sum_{t=1}^{N} (x_t - \mu)^* \Sigma^{-1} (x_t - \mu)\biggr\}. \tag{38.1}$$

Напомним, что отношением правдоподобия для проверки гипотезы H_0 называется статистика

$$\Lambda(X) = \frac{\sup\{L(\mu, \Sigma) \mid (\mu, \Sigma) \in Q\}}{\sup\{L(\mu, \Sigma) \mid (\mu, \Sigma) \in Q_0\}}, \tag{38.2}$$

где

$$Q = \bigl\{(\mu, \Sigma) \bigm| \mu \in \mathbb{R}^n, \quad \Sigma = \Sigma^* > 0\bigr\},$$

$$Q_0 = \bigl\{(\mu, \Sigma) \in Q \bigm| \Sigma_{ij} = 0 \ \text{ при } \ i \neq j\bigr\}.$$

По определению, критерий отношения правдоподобия имеет вид

$$\begin{cases} H_0, & \text{если} \quad \Lambda(X) \leqslant \Delta, \\ H_1, & \text{если} \quad \Lambda(X) > \Delta, \end{cases} \tag{38.3}$$

где критическое значение Δ выбирается таким образом, чтобы вероятность ошибки первого рода $\alpha = P\{\Lambda(X) > \Delta \mid H_0\}$ не превосходила некоторого заранее заданного уровня значимости ε.

Чтобы применять критерий (38.3) для решения практических задач, необходимо в явном виде вычислить $\Lambda(X)$ и Δ.

Вначале найдем приближенное значение константы Δ. Если верна гипотеза независимости H_0, то в тогда силу теоремы 19.1 распределение статистики $2\ln\Lambda(X)$ сходится (при $N \to \infty$) к распределению χ^2_s, где s — коразмерность множества Q_0 в Q. Поэтому при больших N можно считать, что

$$\alpha = P\{\Lambda(X) > \Delta \mid H_0\} = P\{2\ln\Lambda(X) > 2\ln\Delta \mid H_0\} \approx 1 - F_{\chi^2_s}(2\ln\Delta).$$

Приравнивая последнее выражение к ε, получаем $F_{\chi^2_s}(2\ln\Delta) = 1 - \varepsilon$, откуда следует, что

$$\Delta = e^{\delta/2}, \qquad \text{где} \quad \delta = 2\ln\Delta = F^{-1}_{\chi^2_s}(1-\varepsilon). \tag{38.4}$$

Мы доказали, что если определять критическое значение Δ по формулам (38.4), то вероятность ошибки первого рода α будет стремиться к ε при $N \to \infty$. Разность между α и ε можно сделать еще меньше, если использовать уточненные формулы

$$\begin{gathered} \Delta = e^{\delta/2}, \qquad \delta = \frac{N}{g(N)} F^{-1}_{\chi^2_s}(1-\varepsilon), \\ g(N) = N - \frac{3}{2} - \frac{1}{3}\,\frac{n^3 - \sum_{i=1}^k n_i^3}{n^2 - \sum_{i=1}^k n_i^2}. \end{gathered} \tag{38.5}$$

В этом случае можно доказать, что $\alpha - \varepsilon = O(N^{-2})$.

Коразмерность s множества Q_0 в Q равна числу независимых скалярных условий, которыми задается Q_0. Эти условия состоят в том, что элементы матрицы Σ, не входящие ни в какой из диагональных блоков Σ_{ii}, $i = 1, \ldots, k$, должны обращаться в нуль. Всего таких условий $n^2 - \sum_{i=1}^{k} n_i^2$. Но независимых среди них будет вдвое меньше, так как матрица Σ симметрична. Поэтому коразмерность Q_0 в Q равна

$$s = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2\right). \tag{38.6}$$

Для вычисления отношения правдоподобия $\Lambda(X)$ нам потребуется следующая

Лемма 38.1. *Если A — матрица размерности $m\times n$, а B — матрица размерности $n\times m$, то* $\mathrm{tr}(A\times B) = \mathrm{tr}(B\times A)$.

Доказательство. Обозначим через a_{ij} элементы матрицы A, а через b_{ij} элементы матрицы B. Тогда

$$\mathrm{tr}(A\times B) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ji}a_{ij} = \mathrm{tr}(B\times A). \quad \square$$

По выборке $X = (x_1, \ldots, x_N)$ из распределения $\mathcal{N}_n(\mu, \Sigma)$ вычислим выборочное среднее $\bar{x}$ и выборочную матрицу ковариаций $\hat{\Sigma}$:

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t, \qquad \hat{\Sigma} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(x_t - \bar{x})^*.$$

Аналогично μ и Σ, они разбиваются на блоки

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{pmatrix}, \qquad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \ldots & \hat{\Sigma}_{1k} \\ \vdots & & \vdots \\ \hat{\Sigma}_{k1} & \ldots & \hat{\Sigma}_{kk} \end{pmatrix}.$$

В § 27 мы доказали, что $\bar{x}$ и $\hat{\Sigma}$ являются оценками максимального правдоподобия для μ и Σ. Это означает, что

$$\sup_{(\mu,\Sigma)\in Q} L(\mu, \Sigma) = L\bigl(\bar{x}, \hat{\Sigma}\bigr) =$$

$$= \frac{1}{(2\pi)^{nN/2}\bigl|\hat{\Sigma}\bigr|^{N/2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{N}(x_t - \bar{x})^*\hat{\Sigma}^{-1}(x_t - \bar{x})\right\}.$$

В силу леммы 38.1

$$\sum_{t=1}^{N}(x_t-\bar{x})^*\hat{\Sigma}^{-1}(x_t-\bar{x}) = \sum_{t=1}^{N}\operatorname{tr}\Big(\hat{\Sigma}^{-1}(x_t-\bar{x})(x_t-\bar{x})^*\Big) = $$

$$= \operatorname{tr}\Big(\hat{\Sigma}^{-1}N\hat{\Sigma}\Big) = Nn.$$

Значит,

$$\sup_{(\mu,\Sigma)\in Q} L(\mu,\Sigma) \;=\; L\big(\bar{x},\hat{\Sigma}\big) \;=\; \frac{e^{-nN/2}}{(2\pi)^{nN/2}\big|\hat{\Sigma}\big|^{N/2}}. \tag{38.7}$$

Допустим теперь, что $\Sigma_{ij}=0$ при всех $i\neq j$. Тогда блоки координат $x^{(1)}, \ldots, x^{(k)}$ независимы в совокупности, и функция правдоподобия $L(\mu,\Sigma)$ представляется как произведение $\prod_{i=1}^{k} L(\mu_i,\Sigma_{ii})$. Поэтому

$$\sup_{(\mu,\Sigma)\in Q_0} L(\mu,\Sigma) \;=\; \prod_{i=1}^{k}\sup_{\mu_i,\Sigma_{ii}} L(\mu_i,\Sigma_{ii}) \;=\; \prod_{i=1}^{k} L\big(\bar{x}_i,\hat{\Sigma}_{ii}\big).$$

Значения $L\big(\bar{x}_i,\hat{\Sigma}_{ii}\big)$ вычисляются аналогично $L\big(\bar{x},\hat{\Sigma}\big)$ с помощью формулы (38.7). Следовательно,

$$\sup_{(\mu,\Sigma)\in Q_0} L(\mu,\Sigma) \;=\; \prod_{i=1}^{k}\frac{e^{-n_iN/2}}{(2\pi)^{n_iN/2}\big|\hat{\Sigma}_{ii}\big|^{N/2}} \;=\; \frac{e^{-nN/2}}{(2\pi)^{nN/2}}\prod_{i=1}^{k}\frac{1}{\big|\hat{\Sigma}_{ii}\big|^{N/2}}. \tag{38.8}$$

Наконец, подставляя (38.7) и (38.8) в (38.2), получаем искомое выражение для $\Lambda(X)$:

$$\Lambda(X) = \left(\frac{\prod_{i=1}^{k}\big|\hat{\Sigma}_{ii}\big|}{\big|\hat{\Sigma}\big|}\right)^{N/2}. \tag{38.9}$$

Задача 1. Докажите, что в формуле (38.9) выборочные матрицы ковариаций $\hat{\Sigma}_{ii}$ и $\hat{\Sigma}$ можно заменить матрицами выборочных корреляций.

Задача 2. Как выглядят формулы (38.5), (38.6), (38.9) в случае, когда $k=n$ и все блоки $x^{(i)}$ содержат по одной координате случайного вектора x?

Глава 6. ВРЕМЕННЫЕ РЯДЫ

§ 39. Основные понятия

Временные ряды служат математической моделью для различных эволюционных процессов, в которых есть элемент случайности и значения которых наблюдаются в дискретные моменты времени. Типичными примерами временных рядов являются котировки акций на фондовых биржах, ежедневно устанавливаемые обменные курсы валют, разнообразные периодически обновляемые экономические и производственные показатели, ежедневные сводки погоды и т. п. Сам термин «временной ряд» чаще всего употребляется в прикладной статистике, а остальные математики то же самое понятие называют «случайным процессом с дискретным временем».

Формально говоря, *временным рядом* называют любую последовательность случайных величин $x_t = x_t(\omega)$, где ω — элемент вероятностного пространства $(\Omega, \mathfrak{A}, P)$, а индекс t принимает целые или целые неотрицательные значения и имеет смысл дискретных моментов времени. Случайные величины x_t называют *значениями* или *отсчетами* временного ряда в момент времени t, а последовательность $\{x_t(\omega)\}$ при любом фиксированном ω называют *реализацией* или *траекторией* временного ряда.

Ниже для простоты мы будем предполагать, что $x_t \in \mathbb{R}$, хотя все основные понятия и большинство результатов без труда переносятся на многомерный случай $x_t \in \mathbb{R}^n$.

Цилиндром в пространстве реализаций временного ряда называют множество траекторий, удовлетворяющих условиям

$$x_{t_1} \in A_1, \;\; x_{t_2} \in A_2, \;\; \ldots, \;\; x_{t_n} \in A_n,$$

в которых t_k — фиксированные моменты времени, а A_k — произвольные борелевские подмножества в $\mathbb{R}$ (рис. 3). Минимальная σ-алгебра, содержащая все цилиндры, называется *цилиндрической*. Мера цилиндра определяется как число $P\{x_{t_1} \in A_1, \ldots, x_{t_n} \in A_n\}$ (вероятность попадания траектории в цилиндр). Эта мера стандартным образом продолжается на цилиндрическую σ-алгебру. Полученную таким образом меру на цилиндрической σ-алгебре называют распределением вероятностей на пространстве реализаций временного ряда.

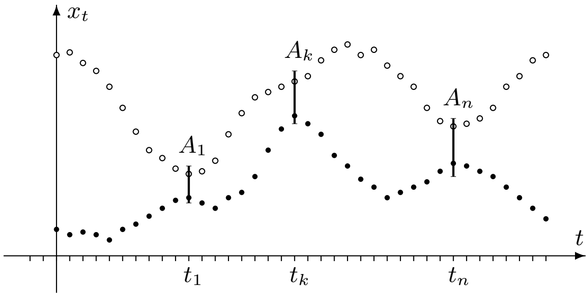


Рис. 3. Цилиндр в пространстве реализаций временного ряда

Задача 1. Докажите, что мера на множестве цилиндров счетно-аддитивна (это необходимо для процедуры продолжения).

Подсказка: прообраз любого цилиндра под действием отображения, сопоставляющего каждому $\omega \in \Omega$ траекторию $\{x_t(\omega)\}$, лежит в $\mathfrak{A}$.

Временной ряд $\{x_t\}$ называют *стационарным в узком смысле*, если он удовлетворяет тождеству

$$P\{x_{t_1} \in A_1, \ldots, x_{t_n} \in A_n\} = P\{x_{t_1+\tau} \in A_1, \ldots, x_{t_n+\tau} \in A_n\} \qquad (39.1)$$

для любых борелевских множеств $A_1, \ldots, A_n$ и натуральных чисел τ. Смысл этого тождества состоит в том, что распределение вероятностей на пространстве реализаций инвариантно относительно сдвигов вдоль оси времени.

Для каждого временного ряда определяют математическое ожидание $\mu_t = \mathsf{E}x_t$, дисперсию $d_t = \mathsf{D}x_t$, а также ковариационную и корреляционную функции:

$$\sigma(t,s) = \mathrm{Cov}\{x_t, x_s\} = \mathsf{E}\{(x_t - \mu_t)(x_s - \mu_s)\},$$

$$\rho(t,s) = \mathrm{Corr}\{x_t, x_s\} = \frac{\sigma(t,s)}{\sqrt{d_t\,d_s}}.$$

Временной ряд $\{x_t\}$ называется *стационарным в широком смысле,* если величины μ_t и d_t конечны и постоянны, а ковариационная функция удовлетворяет тождеству

$$\sigma(t,s) = \sigma(t+\tau, s+\tau).$$

Из тождества (39.1) вытекает, что если временной ряд стационарен в узком смысле и величины μ_t и d_t конечны, то он стационарен также и в широком смысле.

Временной ряд называют *гауссовским*, если все его конечномерные распределения (распределения случайных векторов $(x_{t_1}, \ldots, x_{t_n})$ для любых заданных моментов времени $t_1, \ldots, t_n$) нормальны. Поскольку нормальные распределения полностью определяются своими математическими ожиданиями и ковариациями, для гауссовских временных рядов понятия стационарности в узком и широком смысле совпадают.

Пусть временной ряд $\{x_t\}$ стационарен в широком смысле. Введем для него обозначение $\sigma(\tau) = \sigma(t, t+\tau)$. Очевидно, $\sigma(\tau) = \sigma(-\tau)$. Кроме того, нетрудно видеть, что $|\sigma(\tau)| \leqslant \sigma(0)$. Это вытекает из неравенства Коши — Буняковского:

$$|\sigma(\tau)| = \left|\mathsf{E}\{(x_t - \mu)(x_{t+\tau} - \mu)\}\right| \leqslant \sqrt{\mathsf{E}\{(x_t - \mu)^2\}}\sqrt{\mathsf{E}\{(x_{t+\tau} - \mu)^2\}} = \sigma(0).$$

Спектральной плотностью стационарного в широком смысле временного ряда называется ряд Фурье

$$f(\lambda) = \sum_{\tau=-\infty}^{+\infty} \sigma(\tau) e^{i\lambda\tau} = \sum_{\tau=-\infty}^{+\infty} \sigma(\tau)\cos(\lambda\tau). \qquad (39.2)$$

По признаку Вейерштрасса достаточным условием для его абсолютной сходимости является сходимость ряда $\sum_\tau |\sigma(\tau)|$. Если последний ряд сходится, то спектральная плотность непрерывна и удовлетворяет неравенству $|f(\lambda)| \leqslant \sum_\tau |\sigma(\tau)|$. Очевидно, спектральная плотность четна и 2π-периодична. Если она известна, то значения ковариационной функции $\sigma(\tau)$ могут быть вычислены как коэффициенты Фурье:

$$\sigma(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\lambda) e^{-i\lambda\tau}\, d\lambda = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\lambda)\cos(\lambda\tau)\, d\lambda. \qquad (39.3)$$

Спектральная плотность применяется для обнаружения периодических зависимостей во временном ряде. Равенство (39.3) полезно понимать как разложение ковариационной функции $\sigma(\tau)$ в «континуальную сумму» гармоник $f(\lambda)\cos(\lambda\tau)$. Каждая такая гармоника имеет амплитуду $f(\lambda)$ и период $2\pi/\lambda$. Очевидно, наибольший вклад в сумму вносит гармоника с максимальной амплитудой. Если λ_0 — точка максимума функции $f(\lambda)$, то число $T_0 = 2\pi/\lambda_0$ (иначе говоря, период гармоники с максимальной амплитудой) называется главным периодом временного ряда.

Задача 2. Докажите, что если ряд (39.2) сходится, то $f(\lambda) \geqslant 0$ и

$$f(\lambda) = \lim_{T\to\infty} \frac{1}{2T} \mathsf{E}\left\{\left|\sum_{t=-T}^{T} (x_t - \mu)e^{i\lambda t}\right|^2\right\}.$$

§ 40. Статистические оценки для параметров временных рядов

Статистические оценки для параметров стационарных временных рядов строятся примерно так же, как и для независимых выборок. Но доказывать их состоятельность сложнее, потому что в общем случае значения временного ряда в различные моменты времени зависят друг от друга.

Будем рассматривать стационарный в широком смысле временной ряд $\{x_t\}_{t\in\mathbb{N}}$. По его реализации длительности T мы хотим оценить его математическое ожидание μ, ковариационную функцию $\sigma(\tau)$ и спектральную плотность $f(\lambda)$.

В качестве оценки для μ естественно взять выборочное среднее

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t.$$

Очевидно, эта оценка несмещенная.

Теорема 40.1. *Если временной ряд $\{x_t\}$ стационарен в широком смысле и удовлетворяет условию $\sum_\tau |\sigma(\tau)| < \infty$, то оценка $\bar{x}$ для μ состоятельна, причем*

$$\lim_{T\to\infty} T\,\mathsf{D}\bar{x} = \sum_{\tau=-\infty}^{+\infty} \sigma(\tau). \tag{40.1}$$

Доказательство. Вначале получим равенство (40.1). Для этого проведем выкладку

$$\begin{aligned}
\mathsf{D}\bar{x} &= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathsf{E}\{(x_t-\mu)(x_s-\mu)\} = \\
&= \frac{1}{T^2}\sum_{\tau=0}^{T-1}\sum_{t=1}^{T-\tau}\mathsf{E}\{(x_t-\mu)(x_{t+\tau}-\mu)\} + \\
&\qquad + \frac{1}{T^2}\sum_{\tau=1}^{T-1}\sum_{s=1}^{T-\tau}\mathsf{E}\{(x_{s+\tau}-\mu)(x_s-\mu)\} = \\
&= \frac{1}{T^2}\sum_{\tau=-T+1}^{T-1}(T-|\tau|)\sigma(\tau) = \frac{1}{T}\sum_{|\tau|<T}\left(1-\frac{|\tau|}{T}\right)\sigma(\tau).
\end{aligned}$$

Из нее следует, что при любом $T_0 < T$

$$T\,\mathsf{D}\bar{x} = \sum_{|\tau|<T}\sigma(\tau) - \sum_{|\tau|\leqslant T_0}\frac{|\tau|}{T}\sigma(\tau) - \sum_{T_0<|\tau|<T}\frac{|\tau|}{T}\sigma(\tau). \qquad (40.2)$$

При $T \to +\infty$ первая сумма в правой части (40.2) сходится к $\sum_\tau \sigma(\tau)$, вторая сумма сходится к нулю, а третью можно сделать сколь угодно малой за счет увеличения T_0 и сходимости ряда $\sum_\tau |\sigma(\tau)|$ (проверьте это). Тем самым доказана формула (40.1).

Из нее вытекает, что $\mathsf{D}\bar{x} \to 0$. Тогда в силу неравенства Чебышёва

$$P\big\{|\bar{x}-\mu| \geqslant \varepsilon\big\} \leqslant \frac{\mathsf{D}\bar{x}}{\varepsilon^2} \longrightarrow 0,$$

что по определению означает состоятельность оценки $\bar{x}$. □

В условиях теоремы 40.1 оценка $\bar{x}$ сильно состоятельна, но доказательство этого факта выходит за рамки нашего курса.

Задача 1. Докажите, что для выполнения равенства (40.1) и всей теоремы 40.1 на самом деле достаточно сходимости ряда $\sum_\tau \sigma(\tau)$.

По реализации временного ряда $\{x_t\}$ длительности T строят следующую оценку ковариационной функции:

$$\hat{\sigma}(\tau) = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}(x_t-\bar{x})(x_{t+\tau}-\bar{x}), \qquad 0 \leqslant \tau < T. \qquad (40.3)$$

А если известно математическое ожидание μ, то можно использовать оценку

$$\tilde{\sigma}(\tau) = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}(x_t-\mu)(x_{t+\tau}-\mu), \qquad 0 \leqslant \tau < T. \tag{40.4}$$

Теорема 40.2. *Пусть временной ряд $\{x_t\}$ гауссовский, стационарный и удовлетворяет условию $\sum_t |\sigma(t)| < \infty$. Тогда оценка $\tilde{\sigma}(\tau)$ несмещенная и состоятельная. При этом*

$$\lim_{T\to\infty} T\,\mathsf{D}\{\tilde{\sigma}(\tau)\} = \sum_{t=-\infty}^{+\infty}\big(\sigma^2(t)+\sigma(t-\tau)\sigma(t+\tau)\big). \tag{40.5}$$

Доказательство. Несмещенность оценки $\tilde{\sigma}(\tau)$ следует из того, что математическое ожидание каждого слагаемого в сумме (40.4) по определению равно $\sigma(\tau)$.

Вычислим дисперсию $\tilde{\sigma}(\tau)$. Не ограничивая общности, можно считать, что $\mu = 0$. В этом случае

$$\mathsf{D}\{\tilde{\sigma}(\tau)\} = \frac{1}{(T-\tau)^2}\sum_{t=1}^{T-\tau}\sum_{s=1}^{T-\tau}\mathsf{E}\big\{x_t x_{t+\tau} x_s x_{s+\tau} - \sigma^2(\tau)\big\}.$$

Моменты четвертого порядка для нормальных распределений были вычислены в предложении 22.1. В силу этого предложения

$$\mathsf{E}\{x_t x_{t+\tau} x_s x_{s+\tau}\} = \sigma^2(\tau)+\sigma^2(s-t)+\sigma(s-t-\tau)\sigma(s-t+\tau).$$

Значит,

$$\begin{aligned}
\mathsf{D}\{\tilde{\sigma}(\tau)\} &= \frac{1}{(T-\tau)^2}\sum_{t=1}^{T-\tau}\sum_{s=1}^{T-\tau}\big(\sigma^2(s-t)+\sigma(s-t-\tau)\sigma(s-t+\tau)\big) = \\
&= \frac{1}{(T-\tau)^2}\sum_{l=-T+\tau+1}^{T-\tau-1}(T-\tau-|l|)\big(\sigma^2(l)+\sigma(l-\tau)\sigma(l+\tau)\big) = \\
&= \frac{1}{T-\tau}\sum_{|l|<T-\tau}\left(1-\frac{|l|}{T-\tau}\right)\big(\sigma^2(l)+\sigma(l-\tau)\sigma(l+\tau)\big).
\end{aligned} \tag{40.6}$$

Заметим, что

$$|\sigma(t-\tau)\sigma(t+\tau)| \leqslant C|\sigma(t+\tau)|, \qquad C = \sup_{t\in\mathbb{Z}} |\sigma(t-\tau)|.$$

Поэтому из сходимости ряда $\sum_t |\sigma(t)|$ вытекает абсолютная сходимость ряда в правой части (40.5). Умножая равенство (40.6) на T и заменяя в нем l на t, получаем

$$\begin{aligned} T\,\mathsf{D}\{\tilde\sigma(\tau)\} = \frac{T}{T-\tau} \sum_{|t|<T-\tau} \bigl(\sigma^2(t) + \sigma(t-\tau)\sigma(t+\tau)\bigr) - \\ - \frac{T}{T-\tau} \sum_{|t|\leqslant T_0} \frac{|t|}{T-\tau} \bigl(\sigma^2(t) + \sigma(t-\tau)\sigma(t+\tau)\bigr) - \\ - \frac{T}{T-\tau} \sum_{T_0<|t|<T-\tau} \frac{|t|}{T-\tau} \bigl(\sigma^2(t) + \sigma(t-\tau)\sigma(t+\tau)\bigr). \end{aligned}$$

При $T \to \infty$ первая сумма в правой части последнего равенства сходится к правой части (40.5), вторая сумма сходится к нулю, а третью сумму можно сделать сколь угодно малой за счет увеличения T_0. Тем самым доказана формула (40.5).

Из нее следует, что $\mathsf{D}\{\tilde\sigma(\tau)\}$ стремится к нулю при $T \to \infty$. В силу неравенства Чебышёва (как и в предыдущей теореме) отсюда вытекает состоятельность $\tilde\sigma(\tau)$. $\square$

Теорема 40.3. *Пусть временной ряд $\{x_t\}$ гауссовский, стационарный и удовлетворяет условию $\sum_t |\sigma(t)| < \infty$. Тогда оценка $\hat\sigma(\tau)$ асимптотически несмещенная и состоятельная.*

Задача 2. Докажите эту теорему, рассмотрев разность $\hat\sigma(\tau) - \tilde\sigma(\tau)$.

На самом деле можно доказать, что оценки $\tilde\sigma(\tau)$ и $\hat\sigma(\tau)$ в последних двух теоремах сильно состоятельны.

Спектральную плотность оценивают при помощи *периодограммы*

$$I(\lambda) = \sum_{|\tau|<T} c_\tau e^{i\lambda\tau} = \sum_{|\tau|<T} c_\tau \cos(\lambda\tau), \tag{40.7}$$

где коэффициенты c_τ — это смещенные выборочные ковариации:

$$c_\tau = \frac{1}{T} \sum_{t=1}^{T-|\tau|} (x_t - \bar x)(x_{t+|\tau|} - \bar x) = \frac{T-|\tau|}{T}\,\hat\sigma(\tau).$$

Для вычисления периодограммы можно также использовать формулы

$$I(\lambda) = \frac{1}{T}\big(A^2(\lambda) + B^2(\lambda)\big), \tag{40.8}$$

$$A(\lambda) = \sum_{t=1}^{T}(x_t - \bar{x})\cos(\lambda t), \qquad B(\lambda) = \sum_{t=1}^{T}(x_t - \bar{x})\sin(\lambda t). \tag{40.9}$$

Задача 3. Докажите эквивалентность равенств (40.7) и (40.8).

Задача 4. Объясните, почему в формуле (40.8) требуется меньше вычислений по сравнению с (40.7).

Задача 5. Докажите, что при $\lambda = 2\pi k/T$ величины $A(\lambda)$, $B(\lambda)$ и $I(\lambda)$ не зависят от $\bar{x}$ (иначе говоря, они не изменятся, если в них положить $\bar{x} = 0$). Поэтому выгодно вычислять периодограмму лишь на дискретном наборе частот $\lambda = 2\pi k/T$.

К великому сожалению, оценка $I(\lambda)$ для спектральной плотности несостоятельна. Грубо говоря, это происходит из-за того, что коэффициенты c_τ в сумме (40.7) имеют дисперсии порядка $1/T$, а поскольку число слагаемых равно $2T-1$, дисперсия суммы оказывается порядка единицы и не стремится к нулю при увеличении T. Кроме того, как нетрудно заметить, $A(0) = B(0) = 0$, откуда следует, что $I(0) = 0$, в то время как $f(0) = \sum_\tau \sigma(\tau) \neq 0$.

Состоятельную оценку спектральной плотности строят с помощью процедуры сглаживания. Она состоит в том, что коэффициенты c_τ в сумме (40.7) домножают на специальным образом подобранные множители w_τ, после чего $I(\lambda)$ превращается в оценку

$$\hat{f}(\lambda) = \sum_{|\tau| \leqslant m} w_\tau c_\tau e^{i\lambda\tau}. \tag{40.10}$$

Число $m = m(T)$ подбирают так, чтобы оно росло вместе с T, но при этом отношение $m(T)/T$ сходилось к нулю. Множители w_τ берут вида

$$w_\tau = w\Big(\frac{\tau}{m}\Big),$$

где $w(x)$ — некоторая непрерывная на отрезке $[-1, 1]$ четная функция, удовлетворяющая условию $w(0) = 1$. Например, при $w(x) \equiv 1$ оценка (40.10) называется усеченной выборочной спектральной плотностью, при $w(x) = 1 - |x|$ получается оценка Бартлетта, а при $w(x) = 1 - x^2$ получается оценка Парзена. Если одновременно выполняются условия $m \to \infty$ и $m/T \to 0$ при $T \to \infty$ плюс некоторые дополнительные предположения относительно временного ряда, то оценка (40.10) оказывается состоятельной.

Задача 6. Докажите тождество

$$\hat{f}(\lambda) = \frac{1}{2\pi}\int_{-\pi}^{\pi} I(\lambda - \mu)\varphi(\mu)\,d\mu, \qquad \text{где} \quad \varphi(\mu) = \sum_{|\tau|\leqslant m} w_\tau e^{i\mu\tau}.$$

Оно показывает, что $\hat{f}(\lambda)$ — это результат сглаживания функции $I(\lambda)$ с ядром $\varphi(\lambda)$. В волновой теории ядро $\varphi(\lambda)$ называют *окном пропускания частот.*

Задача 7. Докажите, что для усеченной выборочной спектральной плотности ядро сглаживания совпадает с ядром Дирихле

$$\varphi(\lambda) = \frac{\sin\bigl(m - \frac{1}{2}\bigr)\lambda}{\sin(\lambda/2)},$$

а для оценки Бартлетта оно совпадает с ядром Фейера

$$\varphi(\lambda) = \frac{\sin^2(m\lambda/2)}{m\sin^2(\lambda/2)}.$$

§ 41. Временные ряды авторегрессии

В прикладной статистике большое внимание уделяют прогнозированию значений временного ряда в будущем при условии, что известно его прошлое. С этой целью изучаемый временной ряд стараются представить в виде

$$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, \dots) + \xi_t, \tag{41.1}$$

где $f(x_{t-1}, x_{t-2}, \dots)$ — некоторая неслучайная функция, а ξ_t — случайные возмущения (помехи). Если функция f линейна, а случайные возмущения ξ_t не зависят друг от друга (или, как минимум, некоррелированы), то модель (41.1) называют авторегрессией. Приведем соответствующие формальные определения.

Процесс белого шума (или *слабого белого шума*) — это стационарный временной ряд $\{\xi_t\}$, у которого отсчеты имеют нулевое среднее, постоянную конечную дисперсию и некоррелированы между собой:

$$\mathsf{E}\xi_t = 0, \qquad \mathsf{D}\xi_t = \sigma_\xi^2, \qquad \mathrm{Corr}\{\xi_t, \xi_s\} = \delta_{ts}.$$

Если случайные величины ξ_t независимы в совокупности и одинаково распределены, то белый шум называют *сильным*. Если к тому же они нормально распределены, то белый шум называют гауссовским.

Временной ряд $\{x_t\}$ называется *авторегрессией* порядка p (обозначение $AR(p)$), если для него выполняется равенство

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + c + \xi_t, \tag{41.2}$$

где ξ_t — белый шум, удовлетворяющий условию $\mathrm{Cov}\{\xi_t, x_{t-\tau}\} = 0$ при всех $\tau > 0$. Этот белый шум следует понимать как случайные помехи, вносимые в процесс $\{x_t\}$, которые не зависят от поведения процесса в прошлом.

Характеристическим многочленом авторегрессии (41.2) называют многочлен

$$\alpha(\lambda) = \lambda^p - \alpha_1 \lambda^{p-1} - \ldots - \alpha_p.$$

Изучим простейшие свойства авторегрессии (41.2). Прежде всего заметим, что выражение $y_t = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + c$ есть не что иное, как линейная регрессия случайной величины x_t на совокупность случайных величин $x_{t-1}, \ldots, x_{t-p}$. Это вытекает из того, что по определению авторегрессии для всех $\tau > 0$ выполняются равенства

$$\mathrm{Cov}\{x_t - y_t, x_{t-\tau}\} = \mathrm{Cov}\{\xi_t, x_{t-\tau}\} = 0.$$

Если из уравнения (41.2) исключить белый шум ξ_t и константу c (положить их равными нулю), то оно примет форму

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p}. \tag{41.3}$$

Числовые последовательности, для которых выполняется уравнение (41.3), принято называть *рекуррентными*. Теория рекуррентных последовательностей очень похожа на теорию линейных дифференциальных уравнений с постоянными коэффициентами. Она подробно изложена в приложении, п. IV. В частности, в ней доказывается, что а) последовательности, удовлетворяющие уравнению (41.3), образуют линейное пространство размерности p (при $\alpha_p \neq 0$);

б) базис в этом пространстве составляют последовательности вида

$$x_t = \lambda^t t^n, \qquad n = 0,\, 1,\, \ldots,\, k(\lambda) - 1, \tag{41.4}$$

где λ — любой корень характеристического многочлена, а $k(\lambda)$ — его кратность;

в) все решения неоднородного уравнения (41.2) имеют вид

$$x_t = \sum_{\tau=s}^{t} (c + \xi_\tau) e(t - \tau) + \sum_{\tau=s-p}^{s-1} x_\tau e_{s-\tau}(t - s), \qquad t \geqslant s, \tag{41.5}$$

где s — произвольное целое число, а $e(t)$, $e_1(t)$, $\ldots$, $e_p(t)$ — фиксированные числовые последовательности, удовлетворяющие однородному уравнению (41.3).

Авторегрессия называется *устойчивой*, если все корни ее характеристического многочлена лежат внутри единичного круга на комплексной плоскости. Сформулированные выше утверждения позволяют сделать важные выводы о поведении устойчивой авторегрессии.

Предложение 41.1. *Если выполняется условие устойчивости, то стационарная авторегрессия* (41.2) *существует и имеет вид*

$$x_t = \sum_{\tau=-\infty}^{t} (c + \xi_\tau) e(t - \tau), \tag{41.6}$$

где $e(t)$ *— фиксированное решение однородного уравнения* (41.3).

Доказательство. Если выполняется условие устойчивости, то все последовательности вида (41.4), а также последовательности $e(t)$ и $e_1(t)$, $\ldots$, $e_p(t)$ в (41.5) убывают с экспоненциальной скоростью (при возрастании t). Предположим, что $\{x_t\}$ — стационарная устойчивая авторегрессия. Тогда в формуле (41.5) можно перейти к пределу при $s \to -\infty$, и в пределе получится равенство (41.6).

С другой стороны, временной ряд (41.6) удовлетворяет уравнению авторегрессии, потому что он получается как предел сумм (41.5), если в них взять $x_{s-1} = \ldots = x_{s-p} = 0$. Очевидно, он стационарен в том же смысле, что и белый шум ξ_t. $\square$

Задача 1. Докажите, что условие устойчивости автоматически влечет за собой неравенства $|\alpha_p| < 1$ и $1 - \alpha_1 - \ldots - \alpha_p > 0$.

Задача 2. Докажите, что при возрастании t любая устойчивая авторегрессия сходится (по распределению) к стационарной авторегрессии (41.6).

Задача 3. Докажите, что если хотя бы один корень характеристического многочлена лежит вне единичного круга, то авторегрессия не может быть стационарной.

Рассмотрим стационарную авторегрессию (41.2). Положим $\mu = \mathsf{E}x_t$ и $\sigma(\tau) = \operatorname{Cov}\{x_t, x_{t-\tau}\}$. Вычисляя математическое ожидание от левой и правой частей уравнения (41.2), получаем

$$\mu = \alpha_1\mu + \ldots + \alpha_p\mu + c \quad \Longrightarrow \quad c = \mu(1 - \alpha_1 - \ldots - \alpha_p).$$

Предположим, что $\mu = c = 0$. Умножим уравнение (41.2) на $x_{t-\tau}$ и вычислим математические ожидания от левой и правой частей при $\tau = 0, 1, \ldots, p$. У нас получится система уравнений Юла — Уокера

$$\begin{cases} \sigma(0) = \alpha_1\sigma(1) + \ldots + \alpha_p\sigma(p) + \sigma_\xi^2, & \tau = 0, \\ \sigma(\tau) = \alpha_1\sigma(\tau-1) + \ldots + \alpha_p\sigma(\tau-p), & \tau = 1, \ldots, p. \end{cases} \tag{41.7}$$

Очевидно, при $c \neq 0$ она останется в силе, потому что ковариационная функция $\sigma(\tau)$ не зависит от μ. Если известны значения $\alpha_1, \ldots, \alpha_p$ и σ_ξ^2, то из этой системы можно найти $\sigma(0), \ldots, \sigma(p)$, и наоборот, по известным значениям $\sigma(0), \ldots, \sigma(p)$ можно вычислить коэффициенты авторегрессии и дисперсию белого шума. Если в системе (41.7) заменить истинные значения $\sigma(\tau)$ их выборочными оценками $\hat\sigma(\tau)$, то соответствующие им коэффициенты $\hat\alpha_i$ и $\hat\sigma_\xi^2$ называются оценками Юла — Уокера для α_i и σ_ξ^2.

На самом деле второе из уравнений Юла — Уокера выполняется при всех $\tau > 0$ и позволяет последовательно вычислять все значения $\sigma(\tau)$. Оно с точностью до обозначений совпадает с однородным линейным уравнением (41.3). Поэтому при условии устойчивости последовательность $\sigma(\tau)$, как и всякое решение уравнения (41.3), стремится к нулю с экспоненциальной скоростью.

Для стационарных временных рядов, кроме ковариационной функции $\sigma(\tau)$, рассматривают еще *автокорреляционную функцию* (АКФ)

$$\rho(\tau) = \frac{\operatorname{Cov}\{x_t, x_{t-\tau}\}}{\sqrt{\mathsf{D}\{x_t\}}\sqrt{\mathsf{D}\{x_{t-\tau}\}}} = \frac{\sigma(\tau)}{\sigma(0)},$$

а также *частную автокорреляционную функцию* (ЧАКФ), определяемую как коэффициент частной корреляции между x_t и $x_{t-\tau}$ при фиксированных значениях промежуточных отсчетов $x_{t-1}, \ldots, x_{t-\tau+1}$:

$$r(\tau) = \frac{\mathrm{Cov}\{x_t, x_{t-\tau} | x_{t-1}, \ldots, x_{t-\tau+1}\}}{\sqrt{\mathsf{D}\{x_t | x_{t-1}, \ldots, x_{t-\tau+1}\}} \sqrt{\mathsf{D}\{x_{t-\tau} | x_{t-1}, \ldots, x_{t-\tau+1}\}}}.$$

Напомним, что по определению

$$\mathrm{Cov}\{x_t, x_{t-\tau} | x_{t-1}, \ldots, x_{t-\tau+1}\} = \mathrm{Cov}\{x_t - Px_t,\ x_{t-\tau} - Px_{t-\tau}\},$$

где символ Py обозначает линейную регрессию случайной величины y на набор случайных величин $x_{t-1}, \ldots, x_{t-\tau+1}$.

Если $\{x_t\}$ — устойчивая стационарная авторегрессия, то ее автокорреляционная функция $\rho(\tau)$, как и ковариационная функция $\sigma(\tau)$, экспоненциально убывает. Что касается частной автокорреляционной функции, то имеет место следующее

Предложение 41.2. *Для стационарной авторегрессии порядка p ЧАКФ $r(\tau)$ удовлетворяет равенствам $r(p) = \alpha_p$ и $r(\tau) = 0$ при всех $\tau > p$.*

Доказательство. Не ограничивая общности, можно считать, что $\mathsf{E}x_t = 0$. Тогда уравнение авторегрессии (41.2) принимает вид

$$x_t = \alpha_1 x_{t-1} + \ldots + \alpha_{p-1} x_{t-p+1} + \alpha_p x_{t-p} + \xi_t. \tag{41.8}$$

Обозначим буквой L линейную оболочку случайных величин x_t, $\ldots$, x_{t-p}, буквой l линейную оболочку $x_{t-1}, \ldots, x_{t-p+1}$, а через P — ортогональное проектирование пространства L на подпространство l. По определению, линейные регрессии случайных величин x_t и x_{t-p} на набор $x_{t-1}, \ldots, x_{t-p+1}$ — это их ортогональные проекции Px_t и Px_{t-p} соответственно, а частная корреляция $r(p)$ равна косинусу угла между векторами $x_t - Px_t$ и $x_{t-p} - Px_{t-p}$.

Заметим, что скалярное произведение $(x_t, x_s) = \mathsf{E}\{x_t x_s\} = \sigma(t-s)$ зависит лишь от разности $t - s$. Рассмотрим линейное преобразование пространства L, переводящее векторы x_t, $x_{t-1}, \ldots, x_{t-p+1}$, x_{t-p} соответственно в x_{t-p}, $x_{t-p+1}, \ldots, x_{t-1}$, x_t. Это преобразование сохраняет скалярное произведение и отображает линейное подпространство l в себя. Поэтому разность $x_t - Px_t$ оно переводит в $x_{t-p} - Px_{t-p}$, откуда следует, что $\|x_t - Px_t\| = \|x_{t-p} - Px_{t-p}\|$.

Далее, в силу определения авторегрессии случайная величина ξ_t не коррелирует с $x_{t-1}, \ldots, x_{t-p}$. Значит, $\xi_t \perp l$ и $\xi_t \perp x_{t-p}$. Спроектируем равенство (41.8) на ортогональное дополнение к подпространству l в L. В итоге у нас получится равенство

$$x_t - Px_t = \alpha_p(x_{t-p} - Px_{t-p}) + \xi_t.$$

Из него видно, что векторы $\alpha_p(x_{t-p} - Px_{t-p})$ и ξ_t образуют катеты прямоугольного треугольника с гипотенузой $x_t - Px_t$. Косинус угла между векторами $x_t - Px_t$ и $x_{t-p} - Px_{t-p}$ равен отношению длины катета $\alpha_p(x_{t-p} - Px_{t-p})$ к длине гипотенузы $x_t - Px_t$, то есть α_p. Тем самым мы доказали, что $r(p) = \alpha_p$.

Не ограничивая общности, можно считать, что авторегрессия (41.8) имеет какой угодно порядок $\tau > p$, полагая $\alpha_{p+1}, \ldots, \alpha_\tau$ равными нулю. Поэтому $r(\tau) = \alpha_\tau = 0$. $\square$

§ 42. Оценивание параметров и прогнозирование авторегрессии

Мечтой большинства финансовых аналитиков и брокеров является представление котировок акций на биржах, обменных курсов валют, а также других финансовых и производственных показателей в виде временных рядов авторегрессии. Это позволило бы им с определенной долей достоверности прогнозировать значение моделируемых показателей в будущем и получать за счет этого систематическую прибыль. Хотя здравый смысл подсказывает, что для игроков на бирже (как и для игроков в рулетку) эта цель вряд ли достижима.

Рассмотрим стационарную авторегрессию

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + c + \xi_t. \tag{42.1}$$

Задача состоит в том, чтобы по ее реализации длительности T оценить неизвестные коэффициенты $\alpha_1, \ldots, \alpha_p$, свободный член c и дисперсию возмущений $\sigma_\xi^2 = \mathsf{D}\xi_t$.

Есть как минимум три способа решения этой задачи: с помощью уравнений Юла — Уокера, с помощью метода наименьших квадратов и методом максимального правдоподобия.

При первом способе нужно вначале вычислить выборочное среднее $\bar{x} = \frac{1}{T}(x_1 + \ldots + x_T)$, выборочные ковариации

$$\hat{\sigma}(\tau) = \frac{1}{T - |\tau|} \sum_{t=1}^{T-|\tau|} (x_t - \bar{x})(x_{t+|\tau|} - \bar{x}), \qquad |\tau| \leqslant p,$$

и составить систему уравнений Юла — Уокера

$$\begin{cases} \hat{\sigma}(0) = \hat{\alpha}_1\hat{\sigma}(1) + \ldots + \hat{\alpha}_p\hat{\sigma}(p) + \hat{\sigma}^2_\xi, & \tau = 0, \\ \hat{\sigma}(\tau) = \hat{\alpha}_1\hat{\sigma}(\tau - 1) + \ldots + \hat{\alpha}_p\hat{\sigma}(\tau - p), & \tau = 1, \ldots, p. \end{cases} \tag{42.2}$$

Из этой системы находятся выборочные коэффициенты $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$ и выборочная дисперсия $\hat{\sigma}^2_\xi$. Затем вычисляется оценка

$$\hat{c} = \bar{x}(1 - \hat{\alpha}_1 - \ldots - \hat{\alpha}_p).$$

Заметим, что в силу системы Юла — Уокера величины $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$ и $\hat{\sigma}^2_\xi$ непрерывно зависят от $\hat{\sigma}(0), \ldots, \hat{\sigma}(p)$. Ранее мы доказали, что оценки $\hat{\sigma}(\tau)$ состоятельны. Отсюда вытекает, что оценки $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$ и $\hat{\sigma}^2_\xi$ тоже состоятельны.

В методе наименьших квадратов оценки $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$ и $\hat{c}$ находятся как решение экстремальной задачи

$$RSS = \sum_{t=p+1}^{T} (x_t - \alpha_1 x_{t-1} - \ldots - \alpha_p x_{t-p} - c)^2 \longrightarrow \min, \tag{42.3}$$

а затем оценка $\hat{\sigma}^2_\xi$ вычисляется по формуле $\hat{\sigma}^2_\xi = RSS/(T - p)$. Способ решения этой задачи подробно описан в § 32.

К сожалению, оценки максимального правдоподобия для коэффициентов авторегрессии не удается выписать в явном виде. Их можно находить приближенно разными численными методами.

Можно доказать, что для устойчивой авторегрессии оценки, полученные каждым из перечисленных выше трех методов, состоятельны, и отличаются друг от друга на величину порядка p/T.

Для оценивания порядка авторегрессии оказывается полезной следующая теорема, которую мы приводим без доказательства.

Теорема 42.1. *Если авторегрессия* (42.1) *одновременно устойчивая, стационарная и гауссовская, а коэффициент* α_p *равен нулю, то* *распределение случайной величины* $\sqrt{T}\,\hat{\alpha}_p$ *сходится к стандартному нормальному распределению* $\mathcal{N}(0,1)$ *(где* $\hat{\alpha}_p$ *— оценка Юла — Уокера или МНК-оценка).*

На основе этой теоремы строится критерий для проверки гипотезы $H_0\colon \alpha_p = 0$ против альтернативы $H_1\colon \alpha_p \neq 0$:

$$\begin{cases} \alpha_p = 0, & \text{если } \sqrt{T}\,|\hat{\alpha}_p| \leqslant \Delta, \\ \alpha_p \neq 0, & \text{если } \sqrt{T}\,|\hat{\alpha}_p| > \Delta, \end{cases} \qquad \text{где} \quad \Delta = \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big).$$

С его помощью определяется порядок авторегрессии. Он равен минимальному числу p, при котором оценки $\hat{\alpha}_{p+1}$, $\hat{\alpha}_{p+2}$, … статистически незначимо отличаются от нуля.

Как уже отмечалось, функция $\varphi(x) = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + c$ является линейной регрессией случайной величины x_t на совокупность x_{t-1}, x_{t-2}, … Мы ранее доказали (в предложении 29.1) следующее свойство линейной регрессии: для любой другой линейной функции $\psi(x) = \psi(x_{t-1}, x_{t-2}, \ldots)$ выполняется неравенство

$$\mathsf{E}\big\{(x_t - \varphi(x))^2\big\} \leqslant \mathsf{E}\big\{(x_t - \psi(x))^2\big\}. \tag{42.4}$$

В случае гауссовских временных рядов линейная регрессия $\varphi(x)$ совпадает с условным математическим ожиданием $\mathsf{E}\{x_t \,|\, x_{t-1}, x_{t-2}, \ldots\}$, и потому неравенство (42.4) выполняется для *всех* (борелевских) функций $\psi(x_{t-1}, x_{t-2}, \ldots)$. В силу этих соображений для прогнозирования очередного значения x_t при известных x_{t-1}, x_{t-2}, … целесообразно использовать оценку

$$\hat{x}_t = \hat{\alpha}_1 x_{t-1} + \ldots + \hat{\alpha}_p x_{t-p} + \hat{c}. \tag{42.5}$$

Построенная модель авторегрессии считается адекватной, если последовательность остатков $\hat{\xi}_t = \hat{x}_t - x_t$ распределена как белый шум.

Теорему 42.1 можно вывести из более общего утверждения. Сформулируем его. Пусть $\alpha = (\alpha_1, \ldots, \alpha_p)$, $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)$ и матрица $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ состоит из элементов $\sigma_{ij} = \sigma(i-j)$. Тогда в условиях теоремы 42.1 распределение случайного вектора $\sqrt{T}(\hat{\alpha} - \alpha)$ сходится к $\mathcal{N}_p\big(0, \sigma_\xi^2 \Sigma^{-1}\big)$. В свою очередь, это утверждение несложно доказать с помощью формулы (33.5) для ковариаций коэффициентов линейной регрессии. К сожалению, последняя формула была нами получена для последовательности *независимых* экспериментов в модели линейной регрессии. В случае временных рядов эти эксперименты *зависимы,* и доказать аналог формулы (33.5) значительно труднее.

§ 43. Временные ряды скользящего среднего

Пусть $\{\xi_t\}$ — процесс белого шума. Временной ряд $\{x_t\}$ называют *скользящим средним* порядка q (обозначение $MA(q)$), если он удовлетворяет условию

$$x_t = \xi_t - \beta_1\xi_{t-1} - \beta_2\xi_{t-2} - \ldots - \beta_q\xi_{t-q} + \mu. \tag{43.1}$$

Если же выполняется тождество

$$x_t - \alpha_1 x_{t-1} - \ldots - \alpha_p x_{t-p} = \xi_t - \beta_1\xi_{t-1} - \ldots - \beta_q\xi_{t-q} + c, \tag{43.2}$$

то временной ряд $\{x_t\}$ называется *авторегрессией и скользящим средним* порядков p, q (обозначение $ARMA(p,q)$). Очевидно, по отдельности авторегрессия и скользящее среднее — это частные случаи общей модели $ARMA(p,q)$. Многочлены

$$\alpha(\lambda) = \lambda^p - \alpha_1\lambda^{p-1} - \ldots - \alpha_p, \qquad \beta(\lambda) = \lambda^q - \beta_1\lambda^{q-1} - \ldots - \beta_q$$

называют *характеристическими многочленами* для соответствующих моделей $AR(p)$ и $MA(q)$.

Рассмотрим свойства скользящего среднего (43.1). Из определения вытекает, что оно всегда стационарно (в том же смысле, что и белый шум ξ_t). При этом $\mathsf{E}x_t = \mu$. Умножим уравнение (43.1) на $x_{t-\tau} - \mu$ и проинтегрируем при $\tau = 0, 1, \ldots, q$. У нас получится система

$$\begin{cases} \sigma(0) = (1 + \beta_1^2 + \ldots + \beta_q^2)\sigma_\xi^2, & \tau = 0, \\ \sigma(\tau) = (-\beta_\tau + \beta_1\beta_{\tau+1} + \ldots + \beta_{q-\tau}\beta_q)\sigma_\xi^2, & \tau = 1, \ldots, q. \end{cases} \tag{43.3}$$

Таким образом, мы выразили значения ковариационной функции $\sigma(\tau)$ через коэффициенты $\beta_1, \ldots, \beta_q$ и дисперсию погрешностей σ_ξ^2. Систему (43.3) можно также использовать в обратном направлении: по известным значениям $\sigma(0), \ldots, \sigma(q)$ находить неизвестные $\beta_1, \ldots, \beta_q$ и σ_ξ^2. К сожалению, относительно последних переменных система нелинейна, и ее приходится решать приближенными методами.

Специально отметим, что при $\tau > q$ ковариация $\sigma(\tau)$ и автокорреляционная функция $\rho(\tau) = \sigma(\tau)/\sigma(0)$ обращаются в нуль, потому что в нижнем уравнении системы (43.3) в правой части не будет ни одного слагаемого.

Говорят, что для скользящего среднего выполняется *условие обратимости*, если все корни характеристического многочлена $\beta(\lambda)$ лежат внутри единичного круга на комплексной плоскости.

Сравним уравнения (42.1) и (43.1), определяющие авторегрессию и скользящее среднее. Легко видеть, что они симметричны: они превращаются друг в друга, если поменять местами переменные x_t и ξ_t, а константу c заменить на $-\mu$. При этом условие устойчивости авторегрессии превращается в условие обратимости скользящего среднего и наоборот. Доказанное выше предложение 41.1 означает, что устойчивую авторегрессию можно представить как скользящее среднее бесконечного порядка с экспоненциально убывающими коэффициентами. В силу симметрии делаем вывод, что и обратимое скользящее среднее может быть представлено как авторегрессия бесконечного порядка с экспоненциально убывающими коэффициентами:

$$x_t = \sum_{\tau=1}^{\infty} \eta(\tau) x_{t-\tau} + c + \xi_t, \qquad \eta(\tau) \to 0. \tag{43.4}$$

Предложение 43.1. *У обратимого скользящего среднего частная автокорреляционная функция $r(t)$ убывает с экспоненциальной скоростью.*

Доказательство. Обозначим через $L = \ell(x_1, \ldots, x_{t-1}, 1)$ линейную оболочку случайных величин x_1, …, x_{t-1} и единичной функции, а через P ортогональный проектор гильбертова пространства случайных величин на подпространство L. По определению ЧАКФ $r(t)$ — это косинус угла между векторами $x_t - Px_t$ и $x_0 - Px_0$.

Из формулы (43.1) следует, что $\mathrm{Cov}\{\xi_t, x_{t-\tau}\} = 0$ при всех $\tau > 0$. На геометрическом языке это означает, что в гильбертовом пространстве случайных величин выполняются соотношения

$$\xi_t \perp x_{t-\tau}, \qquad \xi_t \perp L. \tag{43.5}$$

Рассмотрим случайную величину $y_t = \sum_{\tau=t}^{\infty} \eta(\tau) x_{t-\tau}$. Из определения пространства L и формул (43.4), (43.5) вытекает, что

$$x_t - Px_t = y_t - Py_t + \xi_t, \qquad \xi_t \perp (y_t - Py_t). \tag{43.6}$$

Используя (43.5) и (43.6), получаем

$$r(t) = \frac{(x_t - Px_t, x_0 - Px_0)}{\|x_t - Px_t\| \, \|x_0 - Px_0\|} = \frac{(y_t - Py_t, x_0 - Px_0)}{\|x_t - Px_t\| \, \|x_0 - Px_0\|} \leqslant$$

$$\leqslant \frac{\|y_t - Py_t\|}{\|x_t - Px_t\|} \leqslant \frac{\|y_t\|}{\|\xi_t\|}.$$

В последней дроби знаменатель $\|\xi_t\| = \sqrt{\mathsf{D}\xi_t}$ постоянен, а числитель по построению убывает с экспоненциальной скоростью. □

Задача*. Будет ли $r(t)$ стремиться к нулю для любого скользящего среднего?

В теории случайных процессов доказывается, что при довольно широких предположениях каждый стационарный временной ряд можно представить и как авторегрессию, и как скользящее среднее (вообще говоря, бесконечных порядков). Поэтому в приложениях все стационарные временные ряды стараются представить как авторегрессии и скользящие средние (по возможности минимальных порядков).

Поведение АКФ $\rho(\tau)$ и ЧАКФ $r(\tau)$ обычно используют для предварительной идентификации стационарных временных рядов. Если $\rho(\tau)$ убывает с экспоненциальной скоростью, а значения $r(\tau)$ обрываются (становятся очень близки к нулю) при $\tau > p$, то выдвигается гипотеза о том, что мы наблюдаем авторегрессию порядка p. Если $r(\tau)$ убывает с экспоненциальной скоростью, а значения $\rho(\tau)$ обрываются при $\tau > q$, то выдвигается гипотеза о том, что мы наблюдаем скользящее среднее порядка q. Если же АКФ $\rho(\tau)$ и ЧАКФ $r(\tau)$ стремятся к нулю, но не обрываются, то следует предположить, что мы имеем дело с общей моделью $ARMA(p, q)$. В этом случае простого правила для определения порядков p и q нет. Нужно просто перебирать различные пары (p, q) и проверять для них гипотезу о принадлежности рассматриваемого временного ряда к модели $ARMA(p, q)$. Наконец, если какая-либо из функций $\rho(\tau)$, $r(\tau)$ вообще не стремится к нулю, делается вывод о том, что наблюдаемый временной ряд не стационарен.

Изложим вкратце идею, как представить любой стационарный временной ряд $\{x_t\}$ в виде бесконечной авторегрессии или скользящего среднего. Пусть L_t обозначает замкнутое линейное подпространство, порожденное случайными величинами $x_{t-1}, x_{t-2}, \ldots$ и единичной функцией, а P_t — ортогональный проектор на L_t. Определим белый шум ξ_t равенством $\xi_t = x_t - P_t x_t$. По построению случайные величины ξ_t не коррелируют с x_{t-1}, x_{t-2}, … и друг с другом. Авторегрессия будет иметь вид $x_t = P_t x_t + \xi_t$. Осталось только разложить вектор $P_t x_t \in L_t$ в ряд по базису x_{t-1}, x_{t-2}, ... Вообще говоря, это возможно не всегда, и для этого накладываются некоторые дополнительные ограничения на временной ряд. Разложение x_t в бесконечное скользящее среднее будет выглядеть как

$$x_t = (x_t - P_t x_t) + (P_t x_t - P_{t-1} x_t) + (P_{t-1} x_t - P_{t-2} x_t) + \ldots + \mathsf{E} x_t,$$

если только выполняется условие $\lim_{\tau\to\infty} \|P_{t-\tau} x_t - \mathsf{E} x_t\| = 0$.

§ 44. Временные ряды авторегрессии и проинтегрированного скользящего среднего

Самым простым примером авторегрессии и проинтегрированного скользящего среднего служит *процесс случайных блужданий*

$$x_t = x_{t-1} + \xi_t, \tag{44.1}$$

где ξ_t — белый шум. Его можно представлять себе как авторегрессию первого порядка с характеристическим многочленом $\alpha(\lambda) = \lambda - 1$. Его единственный корень равен единице, и поэтому авторегрессия (44.1) неустойчива. Чтобы получить из случайных блужданий устойчивую авторегрессию, рассмотрим временной ряд, состоящий из разностей $\Delta x_t = x_t - x_{t-1}$. Тогда уравнение (44.1) можно записать как $\Delta x_t = \xi_t$. Значит, Δx_t является авторегрессией нулевого порядка (причем устойчивой). Очевидно, для восстановления исходного временного ряда $\{x_t\}$ достаточно знать временной ряд разностей $\{\Delta x_t\}$ и какое-нибудь одно значение x_s:

$$x_t = x_s + \Delta x_{s+1} + \Delta x_{s+2} + \ldots + \Delta x_t. \tag{44.2}$$

Обозначим через B оператор сдвига $Bx_t = x_{t-1}$, а через I — тождественный оператор. Тогда $\Delta x_t = (I - B)x_t$. Разностный оператор порядка d определяется формулами

$$\Delta^d = (I - B)^d; \qquad \Delta^d x_t = \sum_{\tau=0}^{d} (-1)^\tau C_d^\tau x_{t-\tau}.$$

Временной ряд $\{x_t\}$ называется *авторегрессией и проинтегрированным скользящим средним* $ARIMA(p, d, q)$, если он удовлетворяет тождеству

$$\Delta^d x_t - \alpha_1 \Delta^d x_{t-1} - \ldots - \alpha_p \Delta^d x_{t-p} = \xi_t - \beta_1 \xi_{t-1} - \ldots - \beta_q \xi_{t-q}, \tag{44.3}$$

в котором ξ_t — белый шум. Другими словами, разность порядка d от временного ряда $\{x_t\}$ является авторегрессией и скользящим средним порядков p, q.

Определим многочлены

$$\Phi(\lambda) = 1 - \alpha_1\lambda - \ldots - \alpha_p\lambda^p, \qquad \Psi(\lambda) = 1 - \beta_1\lambda - \ldots - \beta_q\lambda^q,$$

двойственные к характеристическим многочленам

$$\alpha(\lambda) = \lambda^p - \alpha_1\lambda^{p-1} - \ldots - \alpha_p, \qquad \beta(\lambda) = \lambda^q - \beta_1\lambda^{q-1} - \ldots - \beta_q.$$

С их помощью равенство (44.3) можно записать так:

$$\Phi(B)(I-B)^d x_t = \Psi(B)\xi_t.$$

Формально временной ряд x_t тоже можно считать авторегрессией и скользящим средним порядков $p+d$ и q, если определять авторегрессию многочленом $\Phi(B)(I-B)^d$. Однако у него есть единичный корень кратности d, и такая авторегрессия неустойчива.

Если у временного ряда не наблюдается систематического тренда, но он при этом не стационарен (нет постоянства дисперсии, или нет экспоненциального стремления к нулю АКФ и ЧАКФ), то его следует попробовать идентифицировать как $ARIMA(p,d,q)$. Для этого нужно последовательно вычислять разности $\Delta^d x_t$ и пытаться идентифицировать их как $ARMA(p,q)$. Если при каком-то (небольшом) d это удается сделать, то можно прогнозировать поведение временного ряда $\Delta^d x_t$, а затем вычислять соответствующий прогноз для исходного ряда x_t, применяя d раз формулу интегрирования (42.2). Построенную модель $ARIMA(p,d,q)$ считают адекватной, если последовательность остатков $\hat{\xi}_t = \hat{x}_t - x_t$ распределена как белый шум (где $\hat{x}_t$ — прогноз для x_t, построенный по x_{t-1}, x_{t-2}, ...).

§ 45. Временные ряды с трендом

Рассмотрим простейшую трендовую модель

$$x_t = f(t) + \xi_t, \tag{45.1}$$

где функция $f(t)$ неслучайна и называется трендом, а ξ_t — белый шум (или хотя бы стационарный временной ряд). Пусть $(x_1, \ldots, x_T)$ — одна реализация этой модели. По ней требуется оценить тренд $f(t)$.

В приложениях принято выделять три типа трендов: систематические (монотонные и медленно изменяющиеся), сезонные (периодические с периодом в один год, одну неделю, одни сутки и т. п.) и комбинированные (результат сложения систематического и сезонного трендов). Обычно тип тренда удается определить визуально по графику временного ряда.

Хорошее приближение к систематическому тренду можно получить с помощью сглаживания исходного временного ряда. Оно заключается в том, что отсчеты x_t заменяются на взвешенные средние

$$\bar{x}_t = \sum_{\tau=-m}^{m} a_\tau x_{t+\tau}, \quad \text{где} \ \ \alpha_\tau \geqslant 0 \ \ \text{и} \sum_{\tau=-m}^{m} a_\tau = 1. \tag{45.2}$$

В силу (45.1) сглаженный временной ряд записывается как

$$\bar{x}_t = \bar{f}(t) + \bar{\xi}_t,$$

где

$$\bar{f}(t) = \sum_{\tau=-m}^{m} a_\tau f(t+\tau) \approx f(t), \qquad \bar{\xi}_t = \sum_{\tau=-m}^{m} a_\tau \xi_{t+\tau}.$$

Для него случайные возмущения $\bar{\xi}_t$ становятся малыми по сравнению с ξ_t, потому что

$$\mathsf{D}\bar{\xi}_t = \sum_{\tau=-m}^{m} a_\tau^2 \mathsf{D}\xi_{t+\tau} \leqslant \max_\tau a_\tau \sum_{\tau=-m}^{m} a_\tau \mathsf{D}\xi_{t+\tau} = \max_\tau a_\tau \mathsf{D}\xi_t.$$

Отсюда видно, что разность между $\bar{x}_t$ и $f(t)$ тоже становится малой.

Сезонный тренд, у которого известен период $T_0 \ll T$, лучше всего оценивать так:

$$\tilde{f}(t) = \frac{1}{n} \sum_{i=0}^{n-1} x_{t+iT_0} \quad (t = 1, \ldots, T_0), \qquad \tilde{f}(t+iT_0) = \tilde{f}(t), \tag{45.3}$$

где $n = [T/T_0]$. Из этого определения и (45.1) вытекает равенство

$$\tilde{f}(t) = f(t) + \frac{1}{n} \sum_{i=0}^{n-1} \xi_{t+iT_0}.$$

Из него видно, что оценка $\tilde{f}(t)$ несмещенная, причем $\mathsf{D}\tilde{f}(t) = \sigma_\xi^2/n$.

Если тренд $f(t)$ комбинированный и известен период его сезонной составляющей, то можно вначале выделить сезонный тренд по формуле (45.3), вычесть его из временного ряда $\{x_t\}$, а затем выделить из разности систематический тренд с помощью сглаживания (45.2).

От сезонного тренда можно также избавиться, перейдя к рассмотрению временного ряда $y_t = x_t - x_{t-T_0}$ (разности с лагом T_0).

После выделения тренда остаточный временной ряд проверяют на стационарность и стараются представить его как белый шум или как авторегрессию и скользящее среднее.

Часто хочется получить не просто набор значений тренда, а какое-нибудь аналитическое выражение для него. Например, это необходимо для прогнозирования временного ряда. Такую задачу обычно решают с помощью разложения тренда по ортогональному базису в пространстве реализаций (метод Галёркина).

Метод Галёркина состоит в том, что реализация временного ряда $(x_1, \ldots, x_T)$ отождествляется с функцией $x(t) = x_t$ и рассматривается как элемент пространства $\mathbb{R}^T$. В этом пространстве выбирают систему ортогональных функций $\varphi_1(t)$, $\varphi_2(t)$, ..., для которых

$$\sum_{t=1}^{T} \varphi_i(t)\varphi_j(t) = 0, \qquad i \neq j.$$

В случае систематического тренда в качестве базисных функций $\varphi_i(t)$ можно взять ортогональные многочлены (полиномы Чебышёва), а для сезонного тренда с периодом T_0 годятся тригонометрические функции $\cos(2\pi it/T_0)$ и $\sin(2\pi it/T_0)$.

Тренд ищется в виде конечной линейной комбинации

$$f(t) = \alpha_1\varphi_1(t) + \ldots + \alpha_m\varphi_m(t).$$

Число m называется *порядком* тренда. Коэффициенты α_i естественно находить с помощью метода наименьших квадратов, как решение экстремальной задачи

$$RSS(\alpha) = \sum_{t=1}^{T}\left(x(t) - \sum_{i=1}^{m}\alpha_i\varphi_i(t)\right)^2 \longrightarrow \min.$$

Решением этой задачи будет ортогональная проекция функции $x(t)$ на линейную оболочку функций $\varphi_1(t)$, ..., $\varphi_m(t)$ (рассматриваемых как элементы пространства $\mathbb{R}^T$). Поэтому оценки для α_i вычисляются как координаты ортогональной проекции:

$$\hat{\alpha}_i = \frac{\sum_{t=1}^T x(t)\varphi_i(t)}{\sum_{t=1}^T \varphi_i^2(t)}.$$

Дисперсию $\sigma_\xi^2 = \mathsf{D}\xi_t$ тоже можно оценить как в модели регрессии:

$$\hat{\sigma}_\xi^2 = \frac{RSS(\hat{\alpha})}{T-m} = \frac{1}{T-m}\sum_{t=1}^T \left(x(t) - \sum_{i=1}^m \hat{\alpha}_i\varphi_i(t)\right)^2.$$

Теорема 45.1. *Если в модели* (45.1) *временной ряд* ξ_t *является белым шумом, а тренд имеет вид* $f(t) = \sum_{i=1}^m \alpha_i\varphi_i(t)$, *то оценки* $\hat{\alpha}_i$ *и* $\hat{\sigma}_\xi^2$ *несмещенные, причем* $\hat{\alpha}_i$ *не коррелированы между собой и*

$$\mathsf{D}\{\hat{\alpha}_i\} = \frac{\sigma_\xi^2}{\sum_{t=1}^T \varphi_i^2(t)}.$$

Если к тому же белый шум ξ_t *гауссовский, то оценки* $\hat{\alpha}_i$ *и* $\hat{\sigma}_\xi^2$ *независимы в совокупности, а статистика*

$$\gamma_i(\alpha_i) = \frac{\hat{\alpha}_i - \alpha_i}{\hat{\sigma}_\xi}\left(\sum_{t=1}^T \varphi_i^2(t)\right)^{1/2} \tag{45.4}$$

имеет распределение Стьюдента с $T-m$ *степенями свободы.*

Доказательство. Все утверждения об $\hat{\alpha}_i$ следуют из выкладок

$$\hat{\alpha}_i = \frac{\sum_{t=1}^T \left(\sum_{j=1}^m \alpha_j\varphi_j(t) + \xi_t\right)\varphi_i(t)}{\sum_{t=1}^T \varphi_i^2(t)} = \alpha_i + \frac{\sum_{t=1}^T \xi_t\varphi_i(t)}{\sum_{t=1}^T \varphi_i^2(t)};$$

$$\mathsf{E}\big\{(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)\big\} = \mathsf{E}\left\{\frac{\sum_{t=1}^T \varphi_i(t)\xi_t}{\sum_{t=1}^T \varphi_i^2(t)}\,\frac{\sum_{s=1}^T \varphi_j(s)\xi_s}{\sum_{s=1}^T \varphi_j^2(s)}\right\} = \frac{\delta_{ij}\sigma_\xi^2}{\sum_{t=1}^T \varphi_i^2(t)}.$$

В случае гауссовского белого шума независимость всех оценок, несмещенность $\hat{\sigma}_\xi^2$ и распределение статистики (45.4) вытекают из леммы о проекциях, доказанной в § 25. $\square$

С помощью статистики (45.4) легко строится критерий для проверки гипотезы $\alpha_m = 0$ против альтернативы $\alpha_m \neq 0$:

$$\begin{cases} \alpha_m = 0, & \text{если} \quad |\gamma_m(0)| \leqslant \Delta, \\ \alpha_m \neq 0, & \text{если} \quad |\gamma_m(0)| > \Delta, \end{cases} \qquad \text{где} \quad \Delta = F^{-1}_{t_{T-m}}\Big(1 - \frac{\alpha}{2}\Big).$$

Этот критерий можно использовать для определения порядка тренда. В идеале порядок m должен быть таким, чтобы коэффициенты $\hat{\alpha}_{m+1}$, $\hat{\alpha}_{m+2}$, ... статистически незначимо отличались от нуля. Однако вместе с увеличением порядка тренд быстро утрачивает гладкость (у его графика возникают узкие «пики» и «впадины»), тогда как по смыслу задачи тренд должен изменяться плавно. Поэтому обычно рассматривают тренды небольших порядков ($m = 1, 2, 3$), стараясь сохранить их гладкость и одновременно обеспечить стационарность остатков.

ПРИЛОЖЕНИЕ

I. Многомерная центральная предельная теорема

В этом подразделе мы будем использовать терминологию и обозначения из § 22.

Теорема. *Если при всех* $j = 1, 2, \ldots$ *случайные векторы* $\xi_j \in \mathbb{R}^m$ *независимы и имеют одинаковые распределения с математическим ожиданием* $\mathsf{E}\xi_j = \mu$ *и матрицей ковариаций* $\mathrm{Cov}\{\xi_j, \xi_j\} = \Sigma$, *то распределение случайного вектора*

$$\zeta_n = \frac{\xi_1 + \ldots + \xi_n - n\mu}{\sqrt{n}}$$

сходится к нормальному распределению $\mathcal{N}_m(0, \Sigma)$.

Это утверждение абсолютно аналогично центральной предельной теореме для скалярных случайных величин. Ниже мы докажем его. Но прежде напомним определение сходимости вероятностных распределений.

Пусть на пространстве $\mathbb{R}^m$ задана последовательность борелевских вероятностных мер $P_1, P_2, \ldots$ Про нее говорят, что она *сходится* (точнее, *слабо сходится*) к борелевской вероятностной мере P, если для любой непрерывной ограниченной функции $f : \mathbb{R}^m \to \mathbb{R}$ имеет место сходимость

$$\int_{\mathbb{R}^m} f(x)\, dP_n(x) \ \longrightarrow\ \int_{\mathbb{R}^m} f(x)\, dP(x).$$

В теории вероятностей доказывается, что сходимость вероятностных мер $P_n \to P$ равносильна сходимости их характеристических функций $\varphi_n(t) \to \varphi(t)$ в каждой точке $t \in \mathbb{R}^m$.

Обозначим через P распределение вероятностей случайных векторов ξ_j из сформулированной выше теоремы. Его характеристическая функция имеет вид

$$\varphi(t) = \mathsf{E}\bigl\{e^{it^*\xi_j}\bigr\} = \int_{\mathbb{R}^m} e^{it^*x}\, dP(x), \qquad t^*x = t_1x_1 + \ldots + t_mx_m.$$

Вычислим частные производные от $\varphi(t)$ первого и второго порядков:

$$\frac{\partial\varphi(t)}{\partial t_k} = \int_{\mathbb{R}^m} ix_k e^{it^*x}\,dP(x), \tag{1}$$

$$\frac{\partial^2\varphi(t)}{\partial t_k \partial t_l} = -\int_{\mathbb{R}^m} x_k x_l e^{it^*x}\,dP(x). \tag{2}$$

Не ограничивая общности, можно считать, что $\mathsf{E}\xi_j = \mu = 0$. Тогда из (1) следует, что $\varphi'(0) = i\mu = 0$, из (2) следует, что $\varphi''(0) = -\Sigma$, и по формуле Тейлора

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \frac{\varphi''(0)}{2}t^2 + o\big(|t|^2\big) = 1 - \frac{1}{2}t^*\Sigma t + \alpha(t)|t|^2,$$

где $\alpha(t) \to 0$ при $t \to 0$.

Вычислим характеристическую функцию случайного вектора ζ_n из центральной предельной теоремы:

$$\varphi_n(t) = \mathsf{E}\big\{e^{it^*\zeta_n}\big\} = \prod_{j=1}^{n} \mathsf{E}\big\{e^{it^*\xi_j/\sqrt{n}}\big\} = \prod_{j=1}^{n} \varphi\bigg(\frac{t}{\sqrt{n}}\bigg) =$$

$$= \bigg(1 - \frac{1}{2n}t^*\Sigma t + \alpha\bigg(\frac{t}{\sqrt{n}}\bigg)\frac{|t|^2}{n}\bigg)^n.$$

Переходя в последнем выражении к пределу, получаем

$$\lim_{n\to\infty} \varphi_n(t) = e^{-\frac{1}{2}t^*\Sigma t}.$$

В правой части этого равенства стоит характеристическая функция нормального закона распределения $\mathcal{N}_m(0, \Sigma)$. Значит, распределение случайного вектора ζ_n сходится к нему.

II. Матрицы Грама

Пусть H — вещественное евклидово или гильбертово пространство со скалярным произведением (x, y). *Матрицей Грама* для набора векторов $x_1, \ldots, x_n \in H$ называется матрица $G = G(x_1, \ldots, x_n)$ размерности $n\times n$, состоящая из элементов $g_{ij} = (x_i, x_j)$.

Изучим простейшие свойства матрицы Грама. Очевидно, она симметрична: $G^* = G$. Кроме того, она неотрицательно определена: для любого вектора $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$

$$\alpha^* G \alpha = \sum_{i=1}^n \sum_{j=1}^n \alpha_i g_{ij} \alpha_j = \sum_{i=1}^n \sum_{j=1}^n (\alpha_i x_i, \alpha_j x_j) = \left| \sum_{i=1}^n \alpha_i x_i \right|^2 \geqslant 0$$

(здесь α понимается как вектор-столбец, а α^* — как вектор-строка).

Предложение 1. *Объем n-мерного параллелепипеда, натянутого на векторы $x_1, \ldots, x_n$, равен $|G|^{1/2}$, а линейная независимость этих векторов равносильна невырожденности матрицы G.*

Доказательство. Обозначим линейную оболочку векторов x_1, …, x_n через $\ell(x_1, \ldots, x_n)$. Очевидно, ее размерность не превосходит n. Поэтому существует ортонормированная система векторов $e_1, \ldots, e_n$, для которой выполняется условие $\ell(x_1, \ldots, x_n) \subset \ell(e_1, \ldots, e_n)$. Разложим каждый из векторов x_i по базису $e_1, \ldots, e_n$:

$$x_i = \sum_{j=1}^n a_{ij} e_j, \qquad i = 1, \ldots, n.$$

Тогда

$$(x_i, x_j) = \sum_{k=1}^n a_{ik} a_{jk}. \tag{1}$$

Составим из элементов a_{ij} матрицу A. Из (1) видно, что $G = AA^*$. Значит, $|G| = |A|^2$. С другой стороны, из линейной алгебры известно, что объем n-мерного параллелепипеда, натянутого на набор векторов $x_1, \ldots, x_n$, равен $\pm|A|$, а независимость этих векторов равносильна невырожденности матрицы A. $\square$

Предложение 2. *Пусть $x_1, \ldots, x_n$, y — векторы из некоторого евклидова пространства, а $P_x y$ обозначает ортогональную проекцию y на линейную оболочку векторов $x_1, \ldots, x_n$. Тогда*

$$|y - P_x y|^2 = \frac{|G(x, y)|}{|G(x)|},$$

где $G(x) = G(x_1, \ldots, x_n)$ и $G(x, y) = G(x_1, \ldots, x_n, y)$.

Доказательство. Это следует из того, что объем параллелепипеда, натянутого на векторы $x_1, \ldots, x_n, y$, равен объему его основания (параллелепипеда, натянутого на векторы $x_1, \ldots, x_n$), помноженному на высоту $|y - P_x y|$. $\square$

Предложение 3. *Если наборы векторов $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ связаны соотношениями*

$$y_i = \sum_{j=1}^{n} a_{ij} x_j, \qquad i = 1, \ldots, m,$$

то для их матриц Грама имеет место равенство $G(y) = AG(x)A^*$, в котором $A = (a_{ij})$.

Доказательство. Это следует из выкладки

$$(y_i, y_j) = \sum_{k=1}^{n} \sum_{l=1}^{n} (a_{ik} x_k, a_{jl} x_l) = \sum_{k=1}^{n} \sum_{l=1}^{n} a_{ik} (x_k, x_l) a_{jl}. \quad \square$$

III. Квадратный корень из неотрицательной симметричной матрицы

Пусть Σ — симметричная неотрицательная матрица размерности $n \times n$ (здесь под неотрицательностью матрицы понимается неотрицательность всех ее собственных значений). В линейной алгебре доказывается, что с помощью ортогонального преобразования ее можно привести к диагональному виду. Другими словами, существует такая ортогональная матрица U, что

$$U \Sigma U^* = \Lambda = \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix}, \qquad UU^* = E,$$

где $\lambda_1, \ldots, \lambda_n$ — собственные значения Σ. Отсюда следует, что сама матрица Σ представляется в виде $\Sigma = U^* \Lambda U$.

Определим квадратный корень из матрицы Λ как

$$\Lambda^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sqrt{\lambda_n} \end{pmatrix}.$$

Положим $\Sigma^{1/2} = U^*\Lambda^{1/2}U$. Тогда матрица $\Sigma^{1/2}$ симметрична, неотрицательна, и для нее

$$\Sigma^{1/2}\Sigma^{1/2} = U^*\Lambda^{1/2}UU^*\Lambda^{1/2}U = U^*\Lambda U = \Sigma.$$

Если все собственные значения λ_i положительны, то у $\Lambda^{1/2}$ существует обратная матрица $\Lambda^{-1/2}$. С ее помощью определяется матрица $\Sigma^{-1/2} = U^*\Lambda^{-1/2}U$, которая будет обратной к $\Sigma^{1/2}$. Действительно,

$$\Sigma^{1/2}\Sigma^{-1/2} = U^*\Lambda^{1/2}UU^*\Lambda^{-1/2}U = U^*U = E.$$

В заключение заметим, что произвольную функцию от симметричной матрицы можно определить при помощи формулы

$$f(\Sigma) = U^*f(\Lambda)U, \qquad \text{где} \quad f(\Lambda) = \begin{pmatrix} f(\lambda_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & f(\lambda_n) \end{pmatrix}.$$

IV. Рекуррентные последовательности

Рекуррентная последовательность порядка n — это последовательность $\{x_t\}_{t\in\mathbb{Z}}$, удовлетворяющая линейному уравнению

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \dots + \alpha_n x_{t-n}, \qquad \alpha_n \neq 0. \tag{1}$$

Теория рекуррентных последовательностей очень напоминает теорию линейных дифференциальных уравнений с постоянными коэффициентами.

Предложение 1. *Множество всех последовательностей, удовлетворяющих уравнению* (1)*, является линейным пространством размерности n.*

Доказательство. Легко видеть, что все решения уравнения (1) образуют линейное пространство. Кроме того, для каждого вектора начальных условий $(x_{-1}, \dots, x_{-n})$ существует только одна последовательность $\{x_t\}_{t\in\mathbb{Z}}$, удовлетворяющая уравнению (1). Действительно, из него по очереди можно вычислить x_0, x_1, x_2, $\dots$, а затем в обратном порядке x_{-n-1}, x_{-n-2}, $\dots$ Тем самым установлен изоморфизм между пространством всех решений уравнения (1) и n-мерным пространством начальных условий $(x_{-1}, \dots, x_{-n})$. $\square$

Характеристическим многочленом для уравнения (1) называется многочлен
$$\alpha(\lambda) = \lambda^n - \alpha_1\lambda^{n-1} - \ \dots \ - \alpha_{n-1}\lambda - \alpha_n.$$
Разложим его на линейные множители:
$$\alpha(\lambda) = (\lambda - \lambda_1)^{k_1} \times \ \dots \ \times (\lambda - \lambda_m)^{k_m}. \tag{2}$$

Предложение 2. *Общее решение уравнения* (1) *имеет вид*
$$x_t = \sum_{i=1}^{m} \lambda_i^t \big(c_{i0} + c_{i1}t + \ \dots \ + c_{i,k_i-1}t^{k_i-1}\big), \tag{3}$$
где λ_i *— все корни многочлена* $\alpha(\lambda)$*, а* k_i *— их кратности.*

Доказательство. Очевидно, уравнение (1) равносильно следующему:
$$x_{t+n} - \alpha_1 x_{t+n-1} - \alpha_2 x_{t+n-2} - \ \dots \ - \alpha_n x_t = 0.$$
Обозначим через B оператор сдвига: $Bx_t = x_{t+1}$. Тогда последнее уравнение можно коротко записать в форме
$$\alpha(B)x_t = 0. \tag{4}$$
При этом в силу (2)
$$\alpha(B) = (B - \lambda_1 I)^{k_1} \dots (B - \lambda_m)^{k_m}, \tag{5}$$
где I — тождественный оператор.

Рассмотрим числовую последовательность $x_t = \lambda_i^t P_k(t)$, в которой $P_k(t)$ — любой многочлен степени k. Для нее
$$(B - \lambda_i I)x_t = \lambda_i^{t+1} P_k(t+1) - \lambda_i^{t+1} P_k(t) = \lambda_i^t P_{k-1}(t),$$
где $P_{k-1}(t)$ — какой-то многочлен степени $k-1$. Отсюда следует, что если $k_i > k$, то имеет место равенство $(B - \lambda_i I)^{k_i} x_t = 0$. Более того, из формулы (5) вытекает, что $\alpha(B)x_t = 0$. Тем самым мы доказали, что любая последовательность вида (3) удовлетворяет уравнению (4). Это уравнение равносильно уравнению (1). Значит, любая последовательность (3) содержится в пространстве решений уравнения (1). Это пространство имеет размерность n. С другой стороны, пространство

последовательностей вида (3) имеет размерность $k_1 + \ldots + k_m = n$, и поэтому совпадает с пространством решений уравнения (1). $\square$

Пусть $\{e(t)\}_{t\in\mathbb{Z}}$ — решение уравнения (1) с начальными условиями

$$e(0) = 1, \quad e(-1) = 0, \quad \ldots, \quad e(-n+1) = 0.$$

Определим последовательность

$$\eta(t) = \begin{cases} e(t), & t \geqslant 0, \\ 0, & t < 0. \end{cases}$$

Она называется *фундаментальным решением* уравнения (1).

Предложение 3. *Для фундаментального решения выполняется уравнение*

$$\eta(t) = \alpha_1\eta(t-1) + \ldots + \alpha_n\eta(t-n) + \delta(t), \tag{6}$$

где $\delta(t)$ *равно единице при* $t = 0$ *и нулю при* $t \neq 0$.

Доказательство. При $t < 0$ левая и правая части уравнения (6) обращаются в нуль. При $t = 0$ это уравнение обращается в истинное равенство $\eta(0) = 1$. Заметим, что по определению $\eta(t) = e(t)$ при всех $t > -n$. Поэтому при положительных t уравнение (6) превращается в истинное равенство $e(t) = \alpha_1 e(t-1) + \ldots + \alpha_n e(t-n)$. $\square$

Наряду с (1) рассмотрим неоднородное линейное уравнение

$$x_t = \alpha_1 x_{t-1} + \ldots + \alpha_n x_{t-n} + \xi_t, \tag{7}$$

в котором последовательность $\{\xi_t\}_{t\in\mathbb{Z}}$ заранее задана. Как и в обычной теории линейных дифференциальных уравнений, общее решение уравнения (7) является суммой любого его частного решения и общего решения однородного уравнения (1). Оказывается, частные решения уравнения (7) легко строятся с помощью фундаментальной последовательности $\eta(t)$.

Введем обозначение

$$Ax_t = \alpha(B)x_t = x_t - \alpha_1 x_{t-1} - \ldots - \alpha_n x_{t-n}.$$

Тогда уравнение (7) принимает вид $Ax_t = \xi_t$, а уравнение (6) можно записать как $A\eta(t) = \delta(t)$. В частности, при любом $\tau \in \mathbb{Z}$ имеет место равенство $A\eta(t-\tau) = \delta(t-\tau)$.

Предложение 4. *Пусть $s \in \mathbb{Z}$. Тогда последовательность*

$$x_t^{(s)} = \sum_{\tau=s}^{+\infty} \xi_\tau \eta(t-\tau) \tag{8}$$

удовлетворяет уравнению (7) *при $t \geqslant s$. А если сходится ряд*

$$x_t^{(\infty)} = \sum_{\tau=-\infty}^{+\infty} \xi_\tau \eta(t-\tau), \qquad t \in \mathbb{Z},$$

то его сумма удовлетворяет уравнению (7) *при всех t.*

Доказательство. Заметим, что в сумме (8) может быть лишь конечное число отличных от нуля слагаемых (поскольку $\eta(t-\tau) = 0$ при $\tau > t$). Применяя к ней оператор A, получаем

$$Ax_t^{(s)} = \sum_{\tau=s}^{+\infty} \xi_\tau A\eta(t-\tau) = \sum_{\tau=s}^{+\infty} \xi_\tau \delta(t-\tau) = \begin{cases} \xi_t, & t \geqslant s, \\ 0, & t < s. \end{cases}$$

Тем самым мы доказали первую часть предложения (4). Вторая часть доказывается точно так же. □

Выберем в пространстве решений однородного уравнения (1) базис, состоящий из последовательностей $e_1(t), \ldots, e_n(t)$, удовлетворяющих начальным условиям

$$e_i(-1) = 0, \ \ldots, \ e_i(-i) = 1, \ \ldots, \ e_i(-n) = 0.$$

Фиксируем любое целое число s. Тогда последовательности $e_1(t-s)$, $\ldots$, $e_n(t-s)$ тоже образуют базис в пространстве решений.

Предложение 5. *Если $\{x_t\}_{t\in\mathbb{Z}}$ — решение уравнения* (7), *то при всех $t \geqslant s$ имеет место равенство*

$$x_t = \sum_{\tau=s}^{t} \xi_\tau e(t-\tau) + \sum_{\tau=s-n}^{s-1} x_\tau e_{s-\tau}(t-s). \tag{9}$$

Доказательство. Рассмотрим две последовательности

$$x_t^{(s)} = \sum_{\tau=s}^{+\infty} \xi_\tau \eta(t-\tau), \qquad y_t^{(s)} = \sum_{\tau=s-n}^{s-1} x_\tau e_{s-\tau}(t-s).$$

В силу предложения 4 последовательность $x_t^{(s)}$ удовлетворяет уравнению (7) при $t \geqslant s$. Кроме того, $x_t^{(s)} = 0$ при $t < s$. Последовательность $y_t^{(s)}$ удовлетворяет однородному уравнению (1) и начальным условиям

$$y_{s-1}^{(s)} = x_{s-1}, \quad \dots, \quad y_{s-n}^{(s)} = x_{s-n}.$$

При фиксированных начальных условиях $x_{s-1}, \dots, x_{s-n}$ есть только одно решение уравнения (7). Поэтому при всех $t \geqslant s$ оно совпадает с суммой $x_t^{(s)} + y_t^{(s)}$. А она совпадает с правой частью (9). $\square$

ЛИТЕРАТУРА

1. *Айвазян, С. А.* Прикладная статистика. Основы эконометрики: в 2 т. / С. А. Айвазян, В. С. Мхитарян. М.: ЮНИТИ-ДАНА, 2001.

2. *Андерсон, Т.* Введение в многомерный статистический анализ / Т. Андерсон. М.: Физматгиз, 1963. 500 с.

3. *Андерсон, Т.* Статистический анализ временных рядов / Т. Андерсон. М.: Мир, 1976. 757 с.

4. *Бокс, Дж.* Анализ временных рядов. Прогноз и управление: в 2 т. / Дж. Бокс, Г. Дженкинс. М.: Мир, 1974. Т. 1. 403 с.

5. *Кремер, Н. Ш.* Эконометрика / Н. Ш. Кремер, Б. А. Путко. М.: ЮНИТИ-ДАНА, 2008. 311 с.

6. *Лагутин, М. Б.* Наглядная математическая статистика / М. Б. Лагутин. М.: Бином, 2007. 472 с.

7. *Харин, Ю. С.* Математическая и прикладная статистика / Ю. С. Харин, Е. Е. Жук. Минск: БГУ, 2005. 279 с.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Учебное издание

Бахтин Виктор Иванович

ВВЕДЕНИЕ В ПРИКЛАДНУЮ СТАТИСТИКУ

КУРС ЛЕКЦИЙ

В двух частях

Часть 2

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ

Ответственный за выпуск *Т. М. Турчиняк*

Дизайн обложки *С. Н. Егоровой*
Технический редактор *Г. М. Романчук*
Корректор *Л. Н. Масловская*

Подписано в печать 24.02.2012. Формат 60×84/16. Бумага офсетная.
Печать офсетная. Усл. печ. л. 5,81. Уч.-изд. л. 4,67.
Тираж 100 экз. Заказ 436.

Белорусский государственный университет.
ЛИ № 02330/0494425 от 08.04.2009.
Пр. Независимости, 4, 220030, Минск.

Республиканское унитарное предприятие
«Издательский центр Белорусского государственного университета».
ЛП № 02330/0494178 от 03.04.2009.
Ул. Красноармейская, 6, 220030, Минск.