Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра информационных технологий
автоматизированных систем

В. С. Муха

Статистические методы обработки данных

Часть 2

Учебно-методическое пособия для студентов специальности "Автоматизированные системы обработки информации"

Минск 2007

6 ТЕОРИЯ СТАТИСТИЧЕСКИХ РЕШЕНИЙ

6.1 Постановка задачи оптимальных статистических решений

Предположим, что исследуемая система может находиться в состоянии s, $s \in S$, S – пространство состояний [11]. Состояние s считается случайной непрерывной величиной с известным распределением $F(s)$. Статистик должен принять решение о состоянии системы (указать значение s), производя или не производя наблюдений (эксперимента) над системой. Решение статистика обозначим $d \in D$, D – пространство решений, такое же, как и пространство состояний S. Решение может быть правильным или неправильным. В случае правильного решения статистик не должен нести потерь, а в случае неправильного решения несет потери. Для характеристики потерь вводится функция потерь $w(s,d)$, зависящая от s и d. Обычно функция потерь зависит от расстояния ρ между s и d. Функция потерь $w(\rho)$ должна удовлетворять следующим свойствам:

1. $w(\rho) \geq 0$ (неотрицательность);
2. $w(0) = 0$ (при $\rho = 0$ потеть нет);
3. если $\rho_1 \geq \rho_2$, то $w(\rho_1) \geq w(\rho_2)$ (неубывание).

Часто для действительных пространств S, D применяется так называемая квадратичная функция потерь

$$w(s,d) = (s-d)^2.$$

Поскольку состояние s системы считается случайным, то функция потерь $w(s,d)$ является случайной величиной.

Теория статистических решений рассматривает так называемые рандомизированные решения, когда решение выносится по жребию, т.е. в соответствии с некоторым законом распределения. В случае рандомизированного решения предполагается, что в пространстве D задано распределение $F(d)$ и решение

выносится как случайное число, выбранное из этого распределения. Функция $F(d)$ называется решающей функцией. В качестве критерия оптимальности решения выбирается средний риск, который представляет собой математическое ожидание функции потерь:

$$r(F(d)) = E(w(s,d)).$$

Задача состоит в определении решающей функции $F(d)$, минимизирующей средний риск $r(F(d))$. Данная оптимизационная задача записывается в виде

$$r(F(d)) \to \min_{F(d)}.$$

Поскольку в этой задаче имеется полная априорная информация в виде функций распределения $F(s)$, $F(d)$, то она относится к классу байесовских задач.

6.2 Статистические решения без наблюдений. Случай непрерывных состояний и решений

Рассмотрим задачу статистических решений, сформулированную в предыдущем разделе, со следующими уточняющими условиями. Предположим, что исследуемая система может находиться в состоянии s, $s \in S$, S – пространство состояний. Состояние s считается случайной непрерывной величиной с известной плотностью вероятности $f(s)$. Статистик должен принять решение $d \in D$, D – пространство решений, *не производя наблюдений над системой*. Будем рассматривать рандомизированные решения, определенные в виде плотности вероятности $f(d)$. Качество решения будем определять средним риском

$$r(f(d)) = E(w(s,d)), \tag{6.1}$$

и решать оптимизационную задачу

$$r(f(d)) \to \min_{f(d)}.$$

Раскроем выражение среднего риска

$$r(f(d)) = E(w(s,d)) = \int_S \int_D w(s,d) f(s) f(d) ds dd .$$

Введем понятие условного риска

$$r(d) = \int_D w(s,d) f(s) ds . \qquad (6.2)$$

Тогда средний риск (6.1) запишется в виде

$$r(f(d)) = \int_D r(d) f(d) dd . \qquad (6.3)$$

В силу неотрицательности функции потерь получаем, что $r(d) \geq 0$. Применяя к (6.3) теорему о среднем значении интегрального исчисления, получим

$$r(f(d)) = \int_D r(d) f(d) dd = r(d)_{cp} \int_D f(d) dd = r(d)_{cp} \geq r(d)_{min} ,$$

где $r(d)_{cp}$ – среднее значение условного риска (6.2), $r(d)_{\min}$ – минимальное значение условного риска (6.2). Последнее равенство в этом выражении записано на основании свойства нормировки для плотности вероятности $\int_D f(d) dd = 1$. Это выражение означает, что средний риск ограничен тем же минимальным значением, что и условный риск. Это значение $r(d)_{\min}$ средний риск будет иметь, если решающую функции выбрать в виде дельта-функции

$$f(d) = \delta(d - d^*) , \qquad (6.4)$$

где d^* – решение, минимизирующее условный риск (6.2) $r(d)$. Действительно, подставив (6.4) в (6.3), мы получим

$$r(f(d)) = \int_D r(d) \delta(d - d^*) dd = r(d^*) = r(d)_{min} .$$

Этот результат мы записали на основе фильтрующего свойства δ-функции: для любой функции $\varphi(x)$ выполняется равенство

$$\int \varphi(x) \delta(x - x^*) dx = \varphi(x^*) .$$

Решающее правило с решающей функцией (6.4) называется нерандомизированным. Итак, мы показали, что оптимальное решающее правило является нерандомизированным и определяется из условия минимума условного риска (6.2):

$$r(d) = \int_S w(s,d) f(s) ds \to \min_{d \in D}.$$

Пример 6.1. Состояние системы s имеет нормальную плотность вероятности

$$f(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-a_s)^2}{2\sigma_s^2}} = N(a_s, \sigma_s^2),$$

где a_s – среднее значение состояния, σ_s^2 – дисперсия состояния. Это значит, что в среднем система находится в состоянии a_s, но может отклоняться от состояния a_s в соответствии с разбросом, который определяется дисперсией σ_s^2. Требуется, не производя наблюдений над системой, указать, в каком состоянии она находится.

Для решения задачи выберем квадратичную функцию потерь

$$w(s,d) = (s-d)^2$$

и найдем условный риск

$$r(d) = \int_{-\infty}^{\infty} w(s,d) f(s) ds = E(s-d)^2 = E(s^2) - 2dE(s) + E(d^2) =$$

$$= E(s^2) - 2dE(s) + d^2 = (a_s^2 + \sigma_s^2) - 2da_s + d^2 \to \min_d.$$

Необходимое условие экстремума функции $r(d)$ есть уравнение

$$\frac{d}{dd} r(d) = 2d - 2a_s = 0,$$

из которого получаем $d^* = a_s$. Следовательно, оптимальным будет решение о том, что система находится в состояние a_s. При этом минимальное значение условного риска равно

$$r(d)_{min} = r(d^*) = a_s^2 + \sigma_s^2 - 2a_s^2 + a_s^2 = \sigma_s^2,$$

т.е. равно дисперсии состояния системы. Приняв решение $d \neq a_s$, мы получим больший риск.

6.3 Статистические решения с наблюдениями. Случай непрерывных состояний и решений

Предположим, что $s \in S$ – состояние системы, S – пространство состояний, и известна плотность вероятности состояния $f(s)$. Над системой выполняется эксперимент, в результате чего получается вектор наблюдений $X = (x_1, x_2, ..., x_n) \in \overline{X}$, $\overline{X}$ – пространство наблюдений. Вектор X, естественно, должен быть каким-то образом связан с состоянием s. Предположим, что эта связь выражается условной плотностью вероятности $f(X/s)$, которую будем считать известной. По этим исходным данным необходимо принять решение $d \in D$, D – пространство решений. Будем считать, что выносимое решение рандомизированное и определяется плотностью вероятности $f(d)$, заданной в пространстве решений D. В данном случае решение зависит от вектора наблюдений X, то есть $d = d(X)$. Качество решения будем характеризовать средним риском

$$r(f(d)) = E(w(s,d)) = \int_S \int_{\overline{X}} \int_D w(s,d) f(d) f(X/s) f(s) ds dX dd, \qquad (6.5)$$

где $w(s,d)$ – функция потерь. Задача состоит в определении решающей функции $f(d(X))$, минимизирующей средний риск:

$$r(f(d(X))) \to \min_{f(d(X))}. \qquad (6.6)$$

Перейдем к решению задачи. Введем условный риск $r(d)$ (при условии, что фиксировано решение d)

$$r(d) = \int_S \int_{\overline{X}} w(s,d) f(s) f(X/s) ds dX. \qquad (6.7)$$

Тогда выражение для среднего риска (6.5) примет вид

$$r(f(d)) = \int_D r(d) f(d) dd.$$

Воспользовавшись применительно к последнему выражению теоремой о среднем значении интегрального исчисления, получим

$$r(f(d)) = r(d)_{cp.} \int_D f(d)dd = r(d)_{cp.} \geq r(d)_{min}.$$

Мы видим, что средний (6.5) и условный (6.7) риски ограничены одним и тем же минимумом. Легко убедиться, что этот минимум достигается при выборе решающей функции вида $f(d) = \delta(d - d^*)$, где $d^* = d^*(X)$ – решение, минимизирующее условный риск $r(d)$ (6.7). Такое решение является нерандомированным. Мы привели задачу к виду

$$r(d) = \int_S \int_{\overline{X}} w(s,d) f(X/s) f(s) dXds \to \min_{d(X)}. \qquad (6.8)$$

Воспользуемся далее теоремой умножения вероятностей $f(X/s)f(s) = f(X)f(s/X)$, где $f(s/X)$ – апостериорная плотность вероятности состояния,

$$f(s/X) = \frac{f(s)f(X/s)}{\int_{-\infty}^{\infty} f(s)f(X/s)ds},$$

и перепишем выражение риска (6.8)

$$r(d) = \int_S \int_{\overline{X}} w(s,d) f(s/X) f(X) dXds.$$

Введем обозначение

$$r(d/X) = \int_S w(s,d) f(s/X) ds.$$

Тогда

$$r(d) = \int_{\overline{X}} r(d/X) f(X) dX.$$

Так как подынтегральное выражение данного функционала не зависит от производных d'_X, то известное из вариационного исчисления уравнение Эйлера для минимизации функционала [10] принимает вид

$$\frac{d}{dd} r(d / X) = 0,$$

т.е. мы получаем следующую оптимизационную задачу:

$$r(d / X) = \int_S w(s,d) f(s / X) ds \to \min_d . \qquad (6.9)$$

Пример 6.2. Состояние системы s описывается нормальной плотностью вероятности

$$f(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-a_s)^2}{2\sigma_s^2}} = N(a_s, \sigma_s^2),$$

где a_s – среднее значение состояния, σ_s^2 – дисперсия состояния. Предположим, что система находится в некотором состоянии из множества возможных состояний, и требуется определить это состояние. Будем считать, что имеется возможность измерять состояние, в котором находится система, однако мы знаем, что это измерение будет содержать ошибку. Пусть измерения описываются плотностью вероятности

$$f(x / s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-s)^2}{2\sigma^2}} = N(s, \sigma^2).$$

Имеется возможность повторить независимые измерения n раз и получить вектор измерений $X = (x_1, ..., x_n)$, плотность вероятности которого

$$f(X / s) = \prod_{i=1}^{n} f(x_i / s) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-s)^2}{2\sigma^2}} .$$

Требуется по этим данным определить состояние системы.

Из изложенной выше теории следует, что оптимальное решение определяется выражением (6.9). Для решения задачи выберем квадратичную функцию потерь $w(s,d) = (s-d)^2$. При квадратичной функции потерь оптимальное решение определяется как апостериорное среднее:

$$d^* = \int_{-\infty}^{\infty} sf(s/X)ds.$$

Заметим, что данная задача аналогична задаче получения байесовской оценки математического ожидания нормального распределения в примере 2.6 раздела 2.5 с очевидной заменой обозначений: a на s, a_0 на a_s и σ_a^2 на σ_s^2. Используя результаты указанного примера, можно записать апостериорную плотность вероятности состояния s в виде

$$f(s/X) = N(d^*, \sigma_{d^*}^2),$$

где $\sigma_{d^*}^2$ – апостериорная дисперсия, d^* – апостериорное среднее, определяемые выражениями (см. пример 2.5 раздела 2.4)

$$\sigma_{d^*}^2 = \frac{1}{\sigma_s^{-2} + n\sigma^{-2}} = \frac{\sigma^2\sigma_s^2}{\sigma^2 + n\sigma_s^2},$$

$$d^* = \frac{1}{\sigma_{d^*}^{-2}}\left(\frac{1}{\sigma_s^2}a_s + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\right).$$

Оптимальным решением является d^*. Минимальное значение риска при этом равно $\sigma_{d^*}^2$. В частном случае отсутствия наблюдений $(n=0)$ мы имеем минимальный риск $\sigma_{d^*}^2 = \sigma_s^2$ и оптимальное решение $d^* = a_s$. Эти же результаты мы получили ранее при рассмотрении задачи без наблюдений в примере 6.1 раздела 6.2.

6.4 Статистические решения с наблюдениями. Случай дискретных состояний, дискретных решений и непрерывных наблюдений

6.4.1 Постановка задачи

Система находится в одном из состояний s_i, $i = \overline{1,L}$, $s_i \in S$. Известны априорные вероятности этих состояний $P(s_i)$, $i = \overline{1,L}$, $\sum_{i=1}^{L} P(s_i) = 1$. Решение $d_i \in D$, $i = \overline{1,L}$, принимается на основе наблюдения над системой $X = (x_1,...,x_n)$, $X \in \overline{X}$. Наблюдения должны быть связаны с состояниями, чтобы предоставлять информацию о них. Будем считать, что эта связь выражается условной плотностью вероятности $f(X / s_i)$, которая предполагается известной. Для характеристики потерь при принятии решения вводится функция потерь $w(s,d)$. Так как состояния и решения дискретны, то эта функция задается в виде матрицы потерь $W = (w_{i,j}) = (w(s_i, d_j))$, $i, j = \overline{1,L}$. Элемент $w_{i,j}$ матрицы потерь характеризует потери, которые несет статистик, когда он принимает решение d_j, в то время как система находится в состоянии s_i (i – номер состояния, j – номер решения). Если $i = j$, то $w_{i,j}$ – потери при правильном решении, поскольку решение совпадает с состоянием. Часто применяется так называемая (0,1)-матрица потерь [20], диагональные элементы которой равны нулю, а остальные – единице. Элементы этой матрицы определяются формулой $w_{i,j} = 1 - \delta_{i,j}$, где $\delta_{i,j} = \begin{Bmatrix} 1, & если & i = j \\ 0, & если & i \neq j \end{Bmatrix}$ – символ Кронекера. Например, при $L = 3$ (0,1)-матрица потерь имеет вид $W = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$. Принимаемое решение d_j зависит от наблюдения X, то есть $d_j = d_j(X)$. Такое решение называется нерандомизированным. В общем случае будем рассматривать рандомизированные решения, то есть решения, принимаемые путем случайного выбора. Это значит, что в пространстве решений D вводится распределение вероятностей $P(d_j)$, $j = \overline{1,L}$, со свойствами

$$0 \le P(d_j) \le 1, \quad \sum_{j=1}^{L} P(d_j) = 1.$$

Качество решения характеризуется средним риском – математическим ожиданием функции потерь $r = E(w(s,d))$, где усреднение производится по всем случайным аргументам функции потерь. Оптимальное решение определяется из условия минимума среднего риска:

$$r(P(d_j)) = E(w(s,d)) \to \min_{P(d_j)}. \qquad (6.10)$$

Средний риск определяется выражением

$$r = r(P(d_j)) = E(w(s,d)) = \int_{\overline{X}} \sum_{i=1}^{L} \sum_{j=1}^{L} w_{i,j} P(s_i) P(d_j(X)) f(X / s_i) dX, \qquad (6.11)$$

где искомые вероятности $P(d_j)$ должны удовлетворять указанным выше свойствам.

6.4.2 Вероятностный смысл среднего риска в случае (0,1)-матрицы потерь

Выясним вероятностный смысл среднего риска (6.11) в случае (0,1)-матрицы потерь. Для этого рассмотрим выражение $P(d_j / s_i) = \int_{\overline{X}} P(d_j(X)) f(X / s_i) dX$, представляющее собой вероятность принятия решения d_j при условии, что система находится в состоянии s_i. С учетом этого обозначения средний риск примет вид:

$$r = \sum_{i=1}^{L} \sum_{j=1}^{L} w_{i,j} P(d_j / s_i) P(s_i).$$

В случае (0,1)-матрицы потерь выражение

$$\sum_{j=1}^{L} w_{i,j} P(d_j / s_i) = \sum_{\substack{j=1 \\ j \ne i}}^{L} P(d_j / s_i) = P(d \ne s / s_i)$$

есть вероятность принятия неверного решения при условии, что система находится в состоянии s_i, то есть условная вероятность принятия неверного решения. Учитывая это выражение, для среднего риска получим

$$r = \sum_{i=1}^{L} P(d \neq s / s_i) P(s_i) = P(d \neq s).$$

Мы видим, что средний риск представляет собой безусловную вероятность принятия неверного решения, или, иначе, безусловную вероятность ошибки $P(d \neq s)$. Следовательно, при (0,1)-матрице потерь оптимальное решение обеспечивает минимальную безусловную вероятность ошибки принятия решения.

6.4.3 Общее решение задачи

Для решения оптимизационной задачи (6.10), в которой средний риск определяется выражением (6.11), введем в рассмотрение условный риск $r(d_j)$

$$r(d_j) = \sum_{i=1}^{L} \int_{\overline{X}} w_{i,j} P(s_i) f(X / s_i) dX . \qquad (6.12)$$

Тогда решаемая задача примет вид

$$r(P(d_j)) = \sum_{j=1}^{L} r(d_j) P(d_j) \to \min_{P(d_j)} .$$

Очевидно, можно воспользоваться дискретным аналогом теоремы интегрального исчисления о среднем значении и записать

$$r(P(d_j)) = r(d_j)_{cp.} \sum_{j=1}^{L} P(d_j) = r(d_j)_{cp} \geq r(d_j)_{min}, \qquad (6.13)$$

где $r(d_j)_{cp.}$ – среднее значение условного риска (6.12), $r(d_j)_{min} = r(d_j^*)$ – минимальное значение условного риска (6.12), d_j^* – решение, минимизирующее условный риск (6.12). Из выражения (6.13) вытекает, что средний риск (6.11) и условный риск (6.12) ограничены одним и тем же минимальным значением

$r(d_j)_{min} = r(d_j^*)$. Это минимальное значение достигается, если выбрать все вероятности $P(d_j) = 0$, кроме одной вероятности $P(d_j^*) = 1$. Такое решение называется нерандомизированным. Итак, оптимальное решение является нерандомизированным и отыскивается из условия минимума условного риска (6.12). Далее, в выражении (6.12) выполним подстановку

$$P(s_i) f(X / s_i) = P(s_i / X) f(X).$$

Тогда получим

$$r(d_j) = \sum_{i=1}^{L} \int_{\overline{X}} w_{i,j} P(s_i / X) f(X) dX, \qquad (6.14)$$

где

$$P(s_i / X) = \frac{P(s_i) f(X / s_i)}{f(X)} - \qquad (6.15)$$

апостериорная вероятность состояния s_i. В выражении (6.14) введем обозначение

$$r(d_j / X) = \sum_{i=1}^{L} w_{i,j} P(s_i / X). \qquad (6.16)$$

Тогда выражение (6.14) будет иметь вид

$$r(d_j) = \int_{\overline{X}} r(d_j / X) f(X) dX. \qquad (6.17)$$

Можно показать, что минимизация риска (6.17) эквивалентна минимизации риска (6.16). Итак, оптимальное решение определяется как решение оптимизационной задачи

$$r(d_j / X) = \sum_{i=1}^{L} w_{i,j} P(s_i / X) \to \min_{d_j}, \qquad (6.18)$$

где $P(s_i / X)$ определяется формулой Байеса (6.15). Подставляя выражение (6.15) в (6.18) с учетом того, что знаменатель $f(X)$ будет присутствовать в каждом слагаемом в (6.18) и не повлияет на минимизацию, получим, что оптимальное решение определяется из условия

$$\psi_j(X) = \sum_{i=1}^{L} w_{i,j} P(s_i) f(X / s_i) \to \min_{d_j}. \qquad (6.19)$$

Для получения оптимального решения необходимо рассчитать значения функций $\psi_j(X)$, $j = \overline{1, L}$, и выбрать решение d_j с таким номером j, которому соответствует минимальное значение указанных функций.

6.4.4 Решение в случае двух состояний

Часто один из многих вариантов выбирают путем парных сравнений вариантов. Поэтому рассмотрим функцию вида

$$\psi_{i,j}(X) = \psi_i(X) - \psi_j(X),\ i = \overline{1, L-1},\ j = \overline{i+1, L},$$

где функции $\psi_i(X)$, $\psi_j(X)$ определяется выражением (6.19). Функция $\psi_{i,j}(X)$ называется дискриминантной или разделяющей. Она призвана выбрать одно из двух решений: d_i или d_j. Решение принимается по следующему правилу:

$$\psi_{i,j}(X) \begin{cases} > 0, \ \textit{решение} \ \ d_j \\ \le 0, \ \ \textit{решение} \ \ d_i \end{cases}. \qquad (6.20)$$

Эта запись означает следующее. Если $\psi_{i,j}(X)$ больше нуля, то принимается решение d_j, в противном случае принимается решение d_i. Такое правило следует из выражения (6.19). Поверхность в n-мерном пространстве измерений $\overline{X}$, определяемая уравнением

$$\psi_{i,j}(X) = 0,$$

называется разделяющей или дискриминантной гиперповерхностью. Решение выносится в зависимости от того, по какую сторону разделяющей гиперповерхности находится вектор наблюдений X.

В случае двух состояний разделяющая функция имеет следующий вид:

$$\psi_{1,2}(X)=\psi_1(X)-\psi_2(X)=w_{1,1}P(s_1)f(X/s_1)+w_{2,1}P(s_2)f(X/s_2)-$$
$$-w_{1,2}P(s_1)f(X/s_1)-w_{2,2}P(s_2)f(X/s_2)=$$
$$=(w_{1,1}-w_{1,2})P(s_1)f(X/s_1)-(w_{2,2}-w_{2,1})P(s_2)f(X/s_2)=$$
$$=c_1f(X/s_1)-c_2f(X/s_2)=f(X/s_2)(c_1l_{1,2}(X)-c_2),$$

где

$$c_2=(w_{2,2}-w_{2,1})P(s_2),$$
$$c_1=(w_{1,1}-w_{1,2})P(s_1),$$
$$l_{1,2}(X)=\frac{f(X/s_1)}{f(X/s_2)}. \qquad (6.21)$$

Функция $l_{1,2}(X)$ называется отношением правдоподобия. Решающее правило (6.20) приобретает следующий вид:

$$c_1l_{1,2}(X)\begin{cases}>c_2,\ решение\ \ d_2\\ \le c_2,\ решение\ \ d_1\end{cases}, \qquad (6.22)$$

Данное решающее правило описывается следующим образом. Вычисляется $c_1l_{1,2}(X)$ и сравнивается с c_2. При $c_1l_{1,2}(X)>c_2$ принимается решение d_2, в противном случае – решение d_1.

Из решающего правила (6.22) можно получить частные случаи. Так, если $(w_{2,2}-w_{2,1})=(w_{1,1}-w_{1,2})<0$, то мы получим правило

$$l_{1,2}(X)\begin{cases}>\dfrac{P(s_2)}{P(s_1)},\ решение\ \ d_1,\\ \le\dfrac{P(s_2)}{P(s_1)},\ решение\ \ d_2.\end{cases}$$

Такое решающее правило известно как правило Зигерта-Котельникова или правило идеального наблюдателя. Если, кроме того, $P(s_1)=P(s_2)$, то мы получим правило

$$l_{1,2}(X)\begin{cases}>1,\ решение\ \ d_1,\\ \le 1,\ решение\ \ d_2.\end{cases}$$

которое известно как правило отношения правдоподобия.

6.4.5 Решение в случае (0,1)-матрицы потерь

В случае (0,1)-матрицы потерь решение упрощается в сторону уменьшения объема вычислений. В этом случае формула (6.19) примет вид

$$\psi_j(X) = \sum_{\substack{i=1 \\ i \neq j}}^{L} P(s_i) f(X / s_i) \to \min_{d_j}.$$

Если учесть, что по формуле полной вероятности $f(X) = \sum_{i=1}^{L} P(s_i) f(X / s_i)$, то получим

$$\psi_j(X) = f(X) - P(s_j) f(X / s_j) \to \min_{d_j}.$$

Последнюю задачу можно записать иначе:

$$\varphi_j(X) = P(s_j) f(X / s_j) \to \max_{j}. \qquad (6.23)$$

Задача (6.23) интерпретируется следующим образом: решение выносится в пользу состояния с таким номером j, для которого величина $P(s_j) f(\bar{x} / s_j)$ наибольшая. Если ввести дискриминантную функцию для двух состояний

$$\varphi_{i,j}(X) = \varphi_i(X) - \varphi_j(X) = P(s_i) f(X / s_i) - P(s_j) f(X / s_j) =$$
$$= f(X / s_j)\left(P(s_i) l_{i,j}(X) - P(s_j)\right),$$

где $l_{i,j}(X)$ – отношение правдоподобия (6.21), то мы получим решающее правило Зигерта-Котельникова

$$l_{i,j}(X) \begin{cases} > \dfrac{P(s_j)}{P(s_i)}, \ \textit{решение} \ \ d_i, \\ \leq \dfrac{P(s_j)}{P(s_i)}, \ \textit{решение} \ \ d_j. \end{cases}$$

где $l_{i,j}(X)$ – отношение правдоподобия (6.21).

Как обычно, можно максимизировать не только функцию $\varphi_j(X)$ (6.23), но и ее логарифм. В этом случае решающее правило приобретает вид

$$ln\, l_{i,j}(X)\begin{cases} > ln\dfrac{P(s_j)}{P(s_i)},\ решение\ \ d_i, \\ \le ln\dfrac{P(s_j)}{P(s_i)},\ \ решение\ \ d_j. \end{cases} \qquad (6.24)$$

6.4.6 Проверка простой двухальтернативной гипотезы о математическом ожидании нормальной генеральной совокупности

Пусть имеется выборка $X=(x_1,x_2,...,x_n)$ из нормальной генеральной совокупности $f_\xi(x/a)=N(a,\sigma^2)$, и требуется проверить гипотезу вида

$$\{H_0: a=a_1; H_1: a=a_2\},$$

где a_1, a_2 – значения дискретной случайной величины, имеющие вероятности $P(a_1)$, $P(a_2)$.

Мы видим, что это задача оптимального статистического решения с двумя состояниями $s_1=a_1$, $s_2=a_2$ и непрерывными наблюдениями. Для решения задачи выберем $(0,1)$-матрицу потерь и воспользуемся логарифмическим отношением правдоподобия

$$ln\, l_{1,2}(X)=ln\frac{f(X/a_1)}{f(X/a_2)}=ln\, f(X/a_1)-ln\, f(X/a_2)$$

и решающим правилом (6.24). В нашем случае

$$f(X/a)=\prod_{i=1}^{n} f_\xi(x_i/a)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} exp\left(-\frac{(x_i-a)^2}{2\sigma^2}\right),$$

$$ln\, f(X/a)=ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n-\sum_{i=1}^{n}\frac{(x_i-a)^2}{2\sigma^2},$$

и мы получим

$$ln\, l_{1,2}(X) = -\sum_{i=1}^{n} \frac{(x_i - a_1)^2}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - a_2)^2}{2\sigma^2} =$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(-x_i^2 + 2x_i a_1 - a_1^2 + x_i^2 - 2x_i a_2 + a_2^2 \right) =$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(2x_i (a_1 - a_2) - (a_1^2 - a_2^2) \right) = n \frac{(a_1 - a_2)}{\sigma^2} \left(\bar{x} - \frac{a_1 + a_2}{2} \right),$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Решающее правило (6.24) для нашей задачи имеет вид

$$n \frac{(a_1 - a_2)}{\sigma^2} \left(\bar{x} - \frac{a_1 + a_2}{2} \right) \begin{cases} > ln\, h, \ \textit{решение} \ \ a = a_1, \\ \leq ln\, h, \ \textit{решение} \ \ a = a_2, \end{cases}$$

где $h = \frac{P(a_2)}{P(a_1)}$ – порог. Предположим, что $a_1 > a_2$. Этого всегда можно достичь выбором гипотезы H_0. Тогда наше решающее правило можно представить в виде

$$\bar{x} \begin{cases} > h_1, \quad \textit{решение} \ \ a = a_1, \\ \leq h_1, \quad \textit{решение} \ \ a = a_2, \end{cases} \qquad (6.25)$$

где $h_1 = \frac{a_1 + a_2}{2} + \frac{\sigma^2 \, ln\, h}{n(a_1 - a_2)}$ – новый порог. Мы видим, что решение о математическом ожидании принимается путем сравнения его оценки $\bar{x}$ с некоторым числом h_1. Это число есть сумма середины отрезка $[a_2, a_1]$ и порога $h_2 = \frac{\sigma^2 \, ln\, h}{n(a_1 - a_2)}$. Геометрический смысл решающего правила (6.25) поясняется рисунком 6.1, на котором изображен отрезок действительной прямой $[a_2, a_1]$ и его середина – точка $c = \frac{a_1 + a_2}{2}$. Решение выносится в пользу правой границы отрезка, если выборочное среднее превышает середину отрезка плюс порог

h_2. Решающее правило имеет особенно простой вид и геометрический смысл в случае $P(a_2) = P(a_1)$. В этом случае $h = 1$, $\ln h = 0$, и мы получаем правило

$$\bar{x}\begin{cases} > \dfrac{a_1 + a_2}{2}, \text{ решение } a = a_1, \\ \text{иначе} - \quad \text{решение } a = a_2, \end{cases},$$

в соответствии с которым решение выносится в пользу того значения a_1 или a_2, ближе к которому находится значение $\bar{x}$. Это значит, что если выборочное среднее $\bar{x}$ попадает в правую половину отрезка $[a_2, a_1]$, то решение выносится в пользу правой границы отрезка, и наоборот.

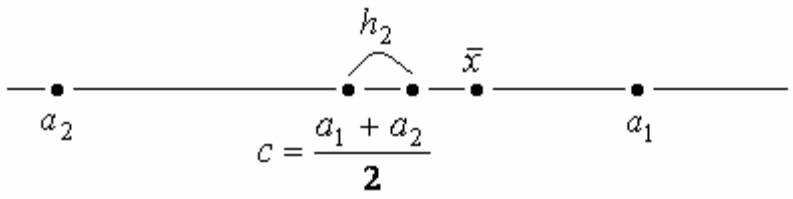

Рис. 6.1. Графическая иллюстрация решающего правила

6.4.7 Статическое распознавание многомерных гауссовских образов

Имеется L образов (классов) $s_1, ..., s_L$. Образы предъявляются на распознавание с вероятностями $\pi_i = P(s_i)$. Если все $\pi_i = 1/L$, то образы поступают на распознавание с одинаковой частотой. Если для некоторого образа вероятность π_j наибольшая, то этот образ s_j будет появляться на распознавание чаще других. Каждый образ характеризуется вектором признаков $X = (x_1, ..., x_n)$, где x_i – отдельный признак. Обычно признаки либо измеряются с ошибками, либо сами являются случайными по своей природе. В любом случае вектор признаков задается известной условной плотностью вероятности $f(X / s_i)$. Требуется по вектору признаков X указать, какому образу он принадлежит. Легко понять, что эта задача является задачей оптимального статического решения с дискрет-

ными состояниями и непрерывными наблюдениями, и ее решение определяется выражением (6.19).

Решим эту задачу для случая (0,1)-матрицы потерь. Решение в этом случае определяется выражением (6.23)

$$\varphi_j(X) = P(s_j) f(X / s_j) \to \max_j, \; j = \overline{1, L},$$

а решающее правило имеет вид (6.24). Пусть вектор признаков распределен по нормальному закону

$$f(X / s_i) = \frac{1}{\sqrt{(2\pi)^n |R_i|}} exp(-\frac{1}{2}(X - A_i)^T R_i^{-1}(X - A_i))),$$

где A_i – n-мерный вектор математического ожидания (вектор истинных значений вектора признаков), R_i – ковариационная (дисперсионная) матрица вектора признаков. В нашем случае

$$ln\, f(X / s_i) = ln\left(\frac{1}{\sqrt{(2\pi)^n |R_i|}}\right) - \frac{1}{2}(X - A_i)^T R_i^{-1}(X - A_i),$$

и логарифмическое отношение правдоподобия (дискриминантная функция) будет иметь вид

$$ln\, l_{i,j}(X) = ln\left(\frac{1}{\sqrt{(2\pi)^n |R_i|}}\right) - \frac{1}{2}(X - A_i)^T R_i^{-1}(X - A_i) -$$

$$- ln\left(\frac{1}{\sqrt{(2\pi)^n |R_j|}}\right) + \frac{1}{2}(X - A_j)^T R_j^{-1}(X - A_j).$$

Мы видим, что в общем случае эта функция является квадратичной функцией вектора признаков X. Дискриминантная кривая в этом случае для двух признаков изображена на рисунке на обложке пособия. Если, $A_i = A_j$, $R_i = R_j$, $P(s_i) = P(s_j)$, т.е. все априорные параметры образов равны между собой, то $ln\, l_{i,j}(X) = 0$ для любого X, и по правилу (6.24) предпочтение всегда будет от-

даваться образу s_j, то есть решающее правило не будет работать. Распознавание может быть успешным, когда образы отличаются хотя бы одним из параметров $P(s_i)$, A_i, R_i. Предположим, что образы отличаются априорными вероятностями $P(s_i)$ и математическими ожиданиями A_i и имеют одинаковые ковариационные матрицы $R_1 = R_2 = ... = R_L = R$. В этом случае логарифмическое отношение правдоподобия имеет вид

$$ln\, l_{i,j}(X) = -\frac{1}{2}(X - A_i)^T R^{-1}(X - A_i) + \frac{1}{2}(X - A_j)^T R^{-1}(X - A_j).$$

Раскрывая скобки и приводя подобные члены, получим

$$ln\, l_{i,j}(X) = -\frac{1}{2}X^T R^{-1}X + \frac{1}{2}X^T R^{-1}X + \frac{1}{2}X^T R^{-1}A_i - \frac{1}{2}X^T R^{-1}A_j +$$

$$+\frac{1}{2}A_i^T R^{-1}X - \frac{1}{2}A_j^T R^{-1}X - \frac{1}{2}A_i^T R^{-1}A_i + \frac{1}{2}A_j^T R^{-1}A_j =$$

$$= X^T R^{-1}(A_i - A_j) - \frac{1}{2}A_i^T R^{-1}A_i + \frac{1}{2}A_j^T R^{-1}A_j.$$

Мы видим, что в этом случае дискриминантная функция является линейной функцией вектора X. Такое решающее правило называется линейным. Дискриминантная гиперповерхность линейного решающего правила является гиперплоскостью в R^n, а в случае двух признаков – прямой на плоскости R^2.

7 ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

7.1 Постановка задачи

Для пояснения сущности дисперсионного анализа вернемся к задаче проверки гипотезы о равенстве математических ожиданий двух нормальных генеральных совокупностей (раздел 5.4.5): имеются две выборки $x_1,..,x_m$ и $y_1,..,y_n$ из двух нормальных генеральных совокупностей $N(a_1,\sigma^2)$ и $N(a_2,\sigma^2)$ с равными, но неизвестными дисперсиями, и нужно проверить двухальтернативную параметрическую сложную гипотезу

$$\{H_0 : a_1 = a_2;\ \ H_1 : a_1 \neq a_2\}.$$

Решение этой задачи в разделе 5.4.5 было основано на сравнении выборочных средних двух генеральных совокупностей и применении t-статистики Стьюдента. В однофакторном дисперсионном анализе проверяется такая же гипотеза, но для произвольного числа генеральных совокупностей. При этом используется иной подход, основанный на сравнении выборочных дисперсий и f-статистики Фишера.

Задача формулируется следующим образом [16]. Имеется k выборок

$$y_{1,1}, y_{1,2},..., y_{1,n_1},\ y_{2,1}, y_{2,2},..., y_{2,n_2},\ \ldots,\ y_{k,1}, y_{k,2},..., y_{k,n_k}$$

разных объемов n_i, $i=\overline{i,k}$, из k нормальных генеральных совокупностей $N(a_1,\sigma^2)$, $N(a_2,\sigma^2)$, , $N(a_k,\sigma^2)$, и нужно проверить гипотезу $H_0 : a_1 = a_2 = ... = a_k$ против альтернативы, что хотя бы одно из этих равенств нарушается.

В краткой форме задача формулируется следующим образом. Имеются наблюдения $y_{i,j}$ вида

$$y_{i,j} = a_i + \xi_{i,j},\ i=\overline{1,k},\ j=\overline{1,n_i}, \qquad (7.1)$$

где $\xi_{i,j} \in N(0,\sigma^2)$ – независимые по i и по j случайные величины. Требуется проверить гипотезу $H_0 : a_1 = a_2 = ... = a_k$ против альтернативы, что не все из этих равенств выполняются.

Эта задача находит следующее практическое применение. Пусть исследуется влияние некоторого фактора на некоторую характеристику производственного процесса, и фактор может принимать k уровней $a_1,...,a_k$. По результатам выполненных экспериментов требуется проверить гипотезу о том, что все уровни равны между собой. Принятие такой гипотезы будет означать, что данный фактор не влияет на данную характеристику. Например, пусть исследуется влияние стажа работы на производительность труда работника. Фактором здесь является стаж работы, а характеристикой – производительность труда. Существуют 4 уровня стажа: до 5 лет, от 5 до 10 лет, от 10 до 15 лет, свыше 15 лет. Выполняются замеры производительности труда работников с различным стажем, по которым проверяется гипотеза $H_0 : a_1 = a_2 = a_3 = a_4$. Принятие этой гипотезы будет означать, что стаж работы не влияет на производительность труда работника.

7.2 Оценки параметров

Для решения задачи, сформулированной в разделе 7.1, нам потребуется найти точечные оценки параметров a_i, σ^2. Воспользуемся для этого методом максимума правдоподобия. Функция правдоподобия наблюдений (7.1) имеет вид

$$L(a_1,...,a_k,\sigma^2) = \prod_{i=1}^{k}\prod_{j=1}^{n_i} f_{\xi,i}(y_{i,j},a_i,\sigma^2) =$$

$$= \prod_{i=1}^{k}\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_{i,j}-a_i)^2\right).$$

Логарифмическая функция правдоподобия будет равна

$$ln\, L(a_1,...,a_k,\sigma^2) \sim \sum_{i=1}^{k}\sum_{j=1}^{n_i}(-\frac{1}{2}ln\,\sigma^2 - \frac{1}{2\sigma^2}(y_{i,j}-a_i)^2), \qquad (7.2)$$

где символ ~ означает эквивалентность для решения задачи. Необходимые условия максимума логарифмической функции правдоподобия представляют собой следующую систему уравнений

$$\frac{\partial}{\partial a_i} ln\, L(a_1,...,a_k,\sigma^2) = \sum_{j=1}^{n_i}\frac{1}{\sigma^2}(y_{i,j}-a_i) = 0,$$

$$\frac{\partial}{\partial \sigma^2} ln\, L(a_1,...,a_k,\sigma^2) = \sum_{i=1}^{k}\sum_{j=1`}^{n_i}(-\frac{1}{2\sigma^2} + \frac{(y_{i,j}-a_i)^2}{2\sigma^4}) = 0,$$

которую можно переписать в виде

$$\sum_{j=1}^{n_i} y_{i,j} - n_i a_i = 0,$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(-\sigma^2 + (y_{i,j}-a_i)^2) = 0.$$

Отсюда получаем

$$\widehat{a}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{i,j},$$

$$\widehat{\sigma}^2_{вн.гр.} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i,j}-\widehat{a}_i)^2,$$

где $n = n_1 + n_2 + ... + n_k$ – общий объем выборки, а оценка дисперсии названа внутригрупповой (*вн.гр.*).

Предположим, что гипотеза H_0 верна, то есть все генеральные совокупности имеют одно и то же (общее) среднее $a_{общ.} = a_1 = a_2 = ... = a_k$, и найдем оценки общего среднего a и общей дисперсии σ^2. Для этого необходимо в логарифмической функции правдоподобия (7.2) все a_i заменить на $a_{общ.}$. В результате получим систему уравнений

$$\frac{\partial}{\partial a} ln\, L = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{1}{\sigma^2} (y_{i,j} - a_{общ.}) = 0,$$

$$\frac{\partial}{\partial \sigma^2} ln\, L = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (-\frac{1}{2\sigma^2} + \frac{(y_{i,j} - a_{общ.})^2}{2\sigma^4}) = 0,$$

из которой находим

$$\hat{a}_{общ.} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{i,j},$$

$$\hat{\sigma}^2_{общ.} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{i,j} - \hat{a}_{общ.})^2.$$

Оценку дисперсии в данном случае мы назвали общей. Оценка $\hat{\sigma}^2_{общ.}$ характеризует разброс наблюдений относительно общего среднего $\hat{a}_{общ.}$.

7.3 Статистика для проверки гипотезы

Рассмотрим вопрос получения статистики для проверки гипотезы дисперсионного анализа. Прибавим и вычтем в скобках в выражении для $\hat{\sigma}^2_{общ.}$ величину $\hat{a}_i$. Получим

$$\hat{\sigma}^2_{общ.} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} ((y_{i,j} - \hat{a}_i) + (\hat{a}_i - \hat{a}_{общ.}))^2 =$$

$$= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} ((y_{i,j} - \hat{a}_i)^2 + 2(y_{i,j} - \hat{a}_i)(\hat{a}_i - \hat{a}_{общ.}) + (\hat{a}_i - \hat{a}_{общ.})^2) =$$

$$= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{i,j} - \hat{a}_i)^2 + \frac{2}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{i,j} - \hat{a}_i)(\hat{a}_i - \hat{a}_{общ.}) + \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\hat{a}_i - \hat{a}_{общ.})^2.$$

Второе слагаемое в последнем выражении равно нулю. Действительно, для этого слагаемого имеем

$$z = \frac{2}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{i,j} - \hat{a}_i)(\hat{a}_i - \hat{a}_{общ.}) = \frac{2}{n} \sum_{i=1}^{k} (\hat{a}_i - \hat{a}_{общ.}) \sum_{j=1}^{n_i} (y_{i,j} - \hat{a}_i). \qquad (7.3)$$

Найдем сумму

$$\sum_{j=1}^{n_i}(y_{i,j}-\hat{a}_i)=\sum_{j=1}^{n_i}y_{i,j}-\sum_{j=1}^{n_i}\hat{a}_i=\sum_{j=1}^{n_i}y_{i,j}-n_i\hat{a}_i=\sum_{j=1}^{n_i}y_{i,j}-\frac{n_i}{n_i}\sum_{j=1}^{n_i}y_{i,j}=0.$$

Поскольку эта сумма является множителем в выражении (7.3) для z, то заключаем, что $z=0$. В результате получаем

$$\hat{\sigma}^2_{общ.}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i,j}-\hat{a}_i)^2+\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{a}_i-\hat{a}_{общ.})^2=\hat{\sigma}^2_{вн.гр.}+\hat{\sigma}^2_{м.гр.}, \qquad (7.4)$$

где введена межгрупповая оценка

$$\hat{\sigma}^2_{м.гр.}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{a}_i-\hat{a}_{общ.})^2=\frac{1}{n}\sum_{i=1}^{k}n_i(\hat{a}_i-\hat{a}_{общ.})^2.$$

Умножив обе части выражения (7.4) на n/σ^2, $n=\sum_{i=1}^{k}n_i$, получим

$$\frac{n\hat{\sigma}^2_{общ.}}{\sigma^2}=\frac{n\hat{\sigma}^2_{вн.гр.}}{\sigma^2}+\frac{n\hat{\sigma}^2_{м.гр.}}{\sigma^2}.$$

Обозначив

$$v_{общ.}=\frac{n\hat{\sigma}^2_{общ.}}{\sigma^2},\ v_{вн.гр.}=\frac{n\hat{\sigma}^2_{вн.гр.}}{\sigma^2},\ v_{м.гр.}=\frac{n\hat{\sigma}^2_{м.гр.}}{\sigma^2},$$

будем иметь

$$v_{общ.}=v_{вн.гр.}+v_{м.гр.}. \qquad (7.5)$$

Поскольку статистика $v_{общ.}$ определяется по n независимым наблюдениям $y_{i,j}$, и при ее определении используется оценка $\hat{a}_{общ.}$ одного параметра, то

$$v_{общ.}\in H_1(n-1).$$

Поскольку статистика $v_{вн.гр.}$ определяется по n независимым наблюдениям $y_{i,j}$, и при ее определении используются оценки k параметров $\hat{a}_1,...,\hat{a}_k$, то

$$v_{вн.гр.}\in H_1(n-k).$$

Из выражения (7.5) по свойству (3.1) распределения хи-квадрат получаем, что

$$v_{мн.гр.} \in H_1(k-1).$$

Детальный анализ показывает, что статистики $v_{общ.}$, $v_{вн.гр.}$, $v_{м.гр.}$ независимы. В этих условиях мы можем записать, что

$$f = \frac{v_{м.гр.}}{k-1} : \frac{v_{вн.гр.}}{n-k} \in F_1(k-1, n-k).$$

Эта статистика и позволяет проверить нашу гипотезу с помощью критерия значимости.

Часто при проверке гипотезы пользуются не оценками дисперсии, а суммами квадратов:

$$s^2_{м.гр.} = \sum_{i=1}^{k} n_i (\hat{a}_i - \hat{a}_{общ.})^2 , \; s^2_{вн.гр.} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{i,j} - \hat{a}_i)^2 .$$

В этом случае статистика для проверки гипотезы определяется выражением

$$f = \frac{s^2_{м.гр.}(n-k)}{s^2_{вн.гр.}(k-1)} \in F_1(k-1, n-k).$$

8 СТАТИСТИКА СЛУЧАЙНЫХ ПРОЦЕССОВ

8.1 Некоторые определения теории случайных процессов

Пусть задано некоторое вероятностное пространство (Ω, F, P).

Случайным процессом называется функция $\xi(\omega, t)$, $\omega \in \Omega$, $t \in \Omega_t$, которая для любого фиксированного t является измеримой функцией аргумента ω.

Аргумент t здесь понимается как время из некоторого промежутка времени Ω_t, а аргумент ω – это элементарный исход (случай). При фиксированном $t = t_1$ мы получаем функцию случая $\xi(\omega, t_1)$, то есть случайную величину, которая называется сечением процесса в момент времени t_1. Если зафиксировать случай $\omega = \omega_1$, то получим функцию времени $\xi(\omega_1, t)$, которая называется реализацией, траекторией или выборочной функцией случайного процесса.

В связи с тем, что чаще всего множество Ω оказывается нам недоступным, то есть элементарные исходы не наблюдаются, случайный процесс обозначают как функцию только времени $\xi(t)$, а зависимость от ω подразумевается.

Если Ω_t – отрезок действительной прямой, то случайный процесс $\xi(t)$ называется процессом с непрерывным временем. Если Ω_t – конечное или счетное множество, то случайный процесс $\xi(t)$ называется случайной последовательностью. Случайная последовательность может быть получена из случайного процесса с непрерывным временем выборкой сечений процесса в дискретные моменты времени.

Конечномерной (n-мерной) функцией распределения случайного процесса $\xi(t)$ называется совместная функция распределения сечений процесса в моменты $t_1, ..., t_n$:

$$F(x_1, ..., x_n, t_1, ..., t_n) = P(\xi(t_1) < x_1, ..., \xi(t_n) < x_n).$$

Конечномерной (n-мерной) плотностью вероятности случайного процесса $\xi(t)$ называется смешанная производная n-го порядка от n-мерной функции распределения:

$$f_\xi(x_1,...,x_n,t_1,...,t_n) = \frac{\partial^n}{\partial x_1,...,\partial x_n} F_\xi(x_1,...,x_n,t_1,...,t_n).$$

В дальнейшем будем рассматривать случайные процессы, для которых существуют конечномерные плотности вероятности.

Математическим ожиданием $E(\xi(t))$ случайного процесса $\xi(t)$ называется функция $a_\xi(t)$, определяемая выражением

$$a_\xi(t) = E(\xi(t)) = \int_{-\infty}^{\infty} x f_\xi(x,t)dx,$$

где $f_\xi(x,t)$ – одномерная плотность вероятности случайного процесса.

Дисперсией $D(\xi(t))$ случайного процесса $\xi(t)$ называется математическое ожидание квадрата отклонения процесса от его математического ожидания:

$$\sigma_\xi^2(t) = D(\xi(t)) = E((\xi(t) - a_\xi(t))^2) = \int_{-\infty}^{\infty} (x - a_\xi(t))^2 f_\xi(x,t)dx.$$

Ковариационной функцией $R_\xi(t_1,t_2)$ случайного процесса $\xi(t)$ называется коэффициент ковариации между сечениями процесса в два момента времени t_1, t_2:

$$R_\xi(t_1,t_2) = cov(\xi(t_1),\xi(t_2)) = E(\overset{\circ}{\xi}(t_1)\overset{\circ}{\xi}(t_2)),\ \overset{\circ}{\xi}(t) = \xi(t) - E(\xi(t)).$$

Для процессов, имеющих конечномерные плотности вероятности, ковариационная функция рассчитывается по формуле

$$R_\xi(t_1,t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 - a_\xi(t_1))(x_2 - a_\xi(t_2))f(x_1,x_2,t_1,t_2)dx_1dx_2,$$

где $f(x_1,x_2,t_1,t_2)$ – двухмерная плотность вероятности случайного процесса.

Случайный процесс $\xi(t)$ называется стационарным в строгом или узком смысле, если его конечномерные распределения инвариантны к сдвигу по оси времени.

Случайный процесс $\xi(t)$ называется стационарным в широком смысле, если его математическое ожидание и дисперсия не зависят от времени, а ковариационная функция зависит от разности своих аргументов:

$$R_\xi(t_1,t_2) = R_\xi(\tau) = E(\overset{\circ}{\xi}(t_1)\overset{\circ}{\xi}(t_1+\tau)),\ \tau = t_1 - t_2.$$

Спектральной плотностью стационарного случайного процесса $\xi(t)$ называется преобразование Фурье от его ковариационной функции $R_\xi(\tau)$:

$$s_\xi(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R_\xi(\tau)e^{-j\omega\tau}d\tau.$$

Тогда ковариационная функция может быть определена как обратное преобразование Фурье от спектральной плотности

$$R_\xi = \int_{-\infty}^{\infty} e^{j\omega\tau}s_\xi(\omega)d\omega.$$

Ковариационная функция стационарной случайной последовательности $\xi(kT)$, $k = 1,2,3...$, T – период квантования во времени, образует последовательность коэффициентов ковариаций $R_\xi(0T)$, $R_\xi(\pm T)$, $R_\xi(\pm 2T)$, ..., $R_\xi(\pm iT),\ldots$, где $R_\xi(iT) = E(\overset{\circ}{\xi}(kT)\overset{\circ}{\xi}(kT+iT))$. Спектральной плотностью $s_{\xi,T}(\omega)$ случайной последовательности $\xi(kT)$ называется ряд Фурье, коэффициентами которого являются значения ковариационной функции:

$$s_{\xi,T}(\omega) = \frac{T}{2\pi}\sum_{k=-\infty}^{\infty} R_\xi(kT)e^{-jk\omega T},\ -\frac{\pi}{T} \le \omega \le \frac{\pi}{T}.$$

Коэффициенты ряда Фурье определяются формулой

$$R_\xi(kT) = \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} e^{jk\omega T} s_{\xi,T}(\omega) d\omega .$$

Это ряд Фурье в комплексной форме. Как известно, функция, имеющая представление в виде ряда Фурье, является периодической, то есть спектральную плотность случайной последовательности достаточно рассматривать в промежутке $-\frac{\pi}{T} < \omega < \frac{\pi}{T}$. Если воспользоваться формулой Эйлера $e^{-jk\omega T} = cos(k\omega T) - j\, sin(k\omega T)$, то спектральную плотность можно записать в виде

$$s_{\xi,T}(\omega) = \frac{T}{2\pi} \sum_{k=-\infty}^{\infty} R_\xi(kT) cos(k\omega T) - j \frac{T}{2\pi} \sum_{k=-\infty}^{\infty} R_\xi(kT) sin(k\omega T) .$$

Если учесть, что синус – функция нечетная ($sin(k\omega T) = -sin(-k\omega T)$), а $R_\xi(kT)$ – четная ($R_\xi(kT) = R_\xi(-kT)$), то второе слагаемое в последнем выражении оказывается равным нулю, и мы получим

$$s_{\xi,T}(\omega) = \frac{T}{2\pi} \sum_{k=-\infty}^{\infty} R_\xi(kT) cos(k\omega T) .$$

Таким образом, спектральная плотность случайной последовательности представляет собой ряд Фурье по косинусам. С учетом четности функции $R_\xi(kT)$ можно записать

$$s_{\xi,T}(\omega) = \frac{T}{2\pi} \left(R_\xi(0) + 2R_\xi(T) cos(\omega T) + 2R_\xi(2T) cos(2\omega T) + \ldots \right) =$$

$$= \frac{T}{\pi} \left(\frac{R_\xi(0)}{2} + R_\xi(T) cos(\omega T) + R_\xi(2T) cos(2\omega T) + \ldots \right).$$

При $\omega = 0$ получим

$$s_{\xi,T}(0) = \frac{T}{\pi} \left(\frac{R_\xi(0)}{2} + R_\xi(T) + R_\xi(2T) + R_\xi(3T) + \ldots \right). \qquad (8.1)$$

8.2 Оценивание математического ожидания стационарной случайной последовательности

Пусть имеем стационарную случайную последовательность $\xi(iT)$, $i = 0,\pm1,\pm2,...$, с математическим ожиданием a_ξ, ковариационной функцией $R_\xi(kT)$, $k = 0,\pm1,\pm2,...$, спектральной плотностью $s_{\xi,T}(\omega)$, T – период квантования во времени, и требуется по реализации этого процесса $x_i = x(iT)$, $i = 1,2,...,n$, найти оценку $\hat{a}_\xi$ математического ожидания a_ξ [1].

В качестве оценки математического ожидания возьмем среднее арифметическое выборочных значений

$$\hat{a}_\xi = \frac{1}{n}\sum_{i=1}^{n} x(iT). \tag{8.2}$$

Для исследования несмещенности данной оценки найдем ее математическое ожидание

$$E(\hat{a}_\xi) = E(\frac{1}{n}\sum_{i=1}^{n} x(iT)) = \frac{1}{n}\sum_{i=1}^{n} E(x(iT)) = \frac{1}{n}\sum_{i=1}^{n} a_\xi = a_\xi.$$

Мы видим, что эта оценка несмещенная. Для исследования состоятельности найдем дисперсию оценки с учетом того, что выборочные значения $x(T), x(2T),..., x(nT)$ являются коррелированными. Получим

$$D(\hat{a}_\xi) = E((\hat{a}_\xi - E(\hat{a}_\xi)^2) = E((\hat{a}_\xi - a_\xi)^2) = E(\frac{1}{n}\sum_{i=1}^{n}(x(iT) - a_\xi))^2 =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(\overset{\circ}{x}(iT)\overset{\circ}{x}(jT)) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} R_\xi((i-j)T).$$

Последняя сумма представляет собой сумму элементов квадратной матрицы. Выполняя суммирование по диагоналям, получим

$$D(\hat{a}_\xi) = \frac{1}{n^2}\left(nR_\xi(0) + 2(n-1)R_\xi(T) + 2(n-2)R_\xi(2T) + ... + 2R_\xi((n-1)T)\right) =$$

$$= \frac{2}{n}\left(\frac{R_\xi(0)}{2} + (1-\frac{1}{n})R_\xi(T) + (1-\frac{2}{n})R_\xi(2T) + ... + \frac{1}{n}R_\xi((n-1)T) \right),$$

$$nD(\hat{a}_\xi) = 2\left(\frac{R_\xi(0)}{2} + (1-\frac{1}{n})R_\xi(T) + (1-\frac{2}{n})R_\xi(2T) + ... + \frac{1}{n}R_\xi((n-1)T) \right).$$

Из данного выражения с учетом выражения (8.1) для $s_{\xi,T}(0)$ мы получаем, что

$$\lim_{n\to\infty} nD(\hat{a}_\xi) = \frac{2\pi}{T} s_{\xi,T}(0),$$

если спектральная плотность $s_{\xi,T}(\omega)$ непрерывна в нуле, то есть если существует $s_{\xi,T}(0)$. Таким образом, оценка (8.2) состоятельна.

8.3 Оценивание ковариационной функции стационарной случайной последовательности

Пусть имеем стационарную случайную последовательность $\xi(iT)$, $i = 0,\pm1,\pm2,...$, с математическим ожиданием a_ξ, ковариационной функцией $R_\xi(kT)$, $k = 0,\pm1,\pm2,...$, спектральной плотностью $s_{\xi,T}(\omega)$, T – период квантования во времени, и требуется по реализации этого процесса $x_i = x(iT)$, $i = 1,2,...,n$, найти оценки $\hat{R}_\xi(kT)$ коэффициентов ковариации $R_\xi(kT)$.

Будем рассматривать два вида оценок ковариационной функции:

1) когда математическое ожидание процесса a_ξ известно;
2) когда математическое ожидание процесса a_ξ неизвестно, а вместо него используется оценка $\hat{a}_\xi$.

8.3.1 Случай известного математического ожидания

Если математическое ожидание известно, то используется оценка вида

$$\hat{R}_\xi(kT) = \frac{1}{n-k}\sum_{i=1}^{n-k} \overset{\circ}{x}(iT)\,\overset{\circ}{x}(iT+kT), \tag{8.3}$$

$$\hat{R}^*_\xi(kT) = \frac{1}{n}\sum_{i=1}^{n-k} \overset{\circ}{x}(iT)\,\overset{\circ}{x}(iT+kT), \tag{8.4}$$

где

$$\overset{\circ}{x}(iT) = x(iT) - a_\xi .$$

Оценки $\hat{R}_\xi(kT)$ и $\hat{R}^*_\xi(kT)$ связаны формулами

$$\hat{R}_\xi(kT) = \frac{n}{n-k}\hat{R}^*_\xi(kT),$$

$$\hat{R}^*_\xi(kT) = \frac{n-k}{n}\hat{R}_\xi(kT) = (1-\frac{k}{n})\hat{R}_\xi(kT) = (1-\frac{|k|}{n})\hat{R}_\xi(kT). \tag{8.5}$$

Возникает вопрос о свойствах этих оценок. Исследуем их несмещенность.

Для оценки $\hat{R}_\xi(kT)$ (8.3) будем иметь

$$E(\hat{R}_\xi(kT)) = \frac{1}{n-k}\sum_{i=1}^{n-k} E(\overset{\circ}{x}(iT)\,\overset{\circ}{x}(iT+kT)) = \frac{1}{n-k}\sum_{i=1}^{n-k} R_\xi(kT) = R_\xi(kT).$$

Мы видим, что оценка $\hat{R}_\xi(kT)$ несмещенная.

Для оценки $\hat{R}^*_\xi(kT)$ (8.4) с учетом формулы (8.5) будем иметь

$$E(\hat{R}^*_\xi(kT)) = (1-\frac{|k|}{n})E(\hat{R}_\xi(kT)) = R_\xi(kT) - \frac{|k|}{n}R_\xi(kT).$$

Видим, что оценка $\hat{R}^*_\xi(kT)$ смещенная, и ее смещение равно $b(R_\xi(kT)) = -\frac{|k|}{n}R_\xi(kT)$. Однако она асимптотически несмещенная, поскольку $b(R_\xi(kT)) \underset{n\to\infty}{\to} 0$.

8.3.2 Случай неизвестного математического ожидания

Если математическое ожидание неизвестно, то используются следующие оценки:

$$\overline{R}_\xi(kT) = \frac{1}{n-k}\sum_{i=1}^{n-k} \overset{*}{x}(iT)\overset{*}{x}(iT+kT), \tag{8.6}$$

$$\overline{R}^*_\xi(kT) = \frac{1}{n}\sum_{i=1}^{n-k} \overset{*}{x}(iT)\overset{*}{x}(iT+kT),$$

где

$$\overset{*}{x}(iT) = x(iT) - \hat{a}_\xi .$$

Оценки $\overline{R}_\xi(kT)$ и $\overline{R}^*_\xi(kT)$ связаны формулами

$$\overline{R}(kT) = \frac{n}{n-k}\overline{R}^*(kT),$$

$$\overline{R}^*_\xi(kT) = \frac{n-k}{n}\overline{R}_\xi(kT) = (1-\frac{k}{n})\overline{R}_\xi(kT) = (1-\frac{|k|}{n})\overline{R}_\xi(kT). \tag{8.7}$$

Для исследования этих оценок на несмещенность найдем математическое ожидание оценки $\overline{R}_\xi(kT)$ (8.6).

$$E(\overline{R}_\xi(kT)) = E\left(\frac{1}{n-k}\sum_{i=1}^{n-k}(x(iT)-\hat{a}_\xi)(x(iT+kT)-\hat{a}_\xi)\right).$$

Прибавим и вычтем a_ξ для каждого множителя данного выражения и обозначим $\overset{\circ}{x}(iT) = x(iT) - a_\xi$. Получим

$$E(\overline{R}_\xi(kT)) = \frac{1}{n-k}\sum_{i=1}^{n-k} E\left(\left(\overset{\circ}{x}(iT) - \frac{1}{n}\sum_{j=1}^{n}\overset{\circ}{x}(jT)\right)\left(\overset{\circ}{x}(iT+kT) - \frac{1}{n}\sum_{m=1}^{n}\overset{\circ}{x}(mT)\right)\right) =$$

$$= \frac{1}{n-k}\sum_{i=1}^{n-k} E\left(\overset{\circ}{x}(iT)\overset{\circ}{x}(iT+kT)\right) - \frac{1}{n(n-k)}\sum_{i=1}^{n-k}\sum_{m=1}^{n} E\left(\overset{\circ}{x}(iT)\overset{\circ}{x}(mT)\right) -$$

$$-\frac{1}{n(n-k)}\sum_{i=1}^{n-k}\sum_{j=1}^{n}E\left(\overset{\circ}{x}(jT)\overset{\circ}{x}(iT+kT)\right)+\frac{1}{n^2(n-k)}\sum_{i=1}^{n-k}\sum_{j=1}^{n}\sum_{m=1}^{n}E\left(\overset{\circ}{x}(jT)\overset{\circ}{x}(mT)\right)=$$

$$=R_\xi(kT)-\frac{1}{n(n-k)}\sum_{i=1}^{n-k}\sum_{m=1}^{n}R((i-m)T)-\frac{1}{n(n-k)}\sum_{i=1}^{n-k}\sum_{j=1}^{n}R((i+k-j)T)+$$

$$+\frac{1}{n^2}\sum_{j=1}^{n}\sum_{m=1}^{n}R((j-m)T)=R_\xi(kT)+b.$$

Мы видим, что оценка $\overline{R}_\xi(kT)$ (8.6) является смещенной. Можно, однако, показать, что смещение b стремится к нулю при $n\to\infty$, и порядок смещения равен $1/n$. Более конкретно, доказана следующая теорема [1].

Теорема 8.1. Если $\sum_{k=-\infty}^{\infty}R_\xi(kT)<\infty$, то

$$\lim_{n\to\infty} n\left[E(\overline{R}_\xi(kT))-R_\xi(kT)\right]=-\sum_{k=-\infty}^{\infty}R_\xi(kT).$$

Если $s_{\xi,T}(\omega)$ непрерывна при $\omega=0$, то

$$\lim_{n\to\infty} n\left[E(\overline{R}_\xi(kT))-R_\xi(kT)\right]=-\frac{2\pi}{T}s_{\xi,T}(0).$$

8.4 Оценивание спектральной плотности стационарной случайной последовательности

Пусть имеем стационарную случайную последовательность $\xi(iT)$, $i=0,\pm1,\pm2,\ldots$, с математическим ожиданием a_ξ, ковариационной функцией $R_\xi(kT)$, $k=0,\pm1,\pm2,\ldots$, спектральной плотностью $s_{\xi,T}(\omega)$, T – период квантования во времени, и требуется по реализации этого процесса $x_i=x(iT)$, $i=1,2,\ldots,n$, найти оценку $\hat{s}_{\xi,T}(\omega)$ спектральной плотности $s_{\xi,T}(\omega)$ этой последовательности.

Будем, как и в разделе 8.3, различать случаи известного и неизвестного математического ожидания случайной последовательности.

8.4.1 Случай известного математического ожидания

Рассмотрим сначала случай, когда математическое ожидание известно, как наиболее простой для анализа.

Простейшей является оценка вида

$$\widehat{s}_{\xi,T}(\omega)=\frac{nT}{8\pi}(A^2(\omega)+B^2(\omega)),\ -\frac{\pi}{T}\le\omega\le\frac{\pi}{T}, \tag{8.8}$$

где

$$A(\omega)=\frac{2}{n}\sum_{i=1}^{n}\overset{\circ}{x}(iT)\cos(i\omega T),$$

$$B(\omega)=\frac{2}{n}\sum_{i=1}^{n}\overset{\circ}{x}(iT)\sin(i\omega T),$$

$$\overset{\circ}{x}(iT)=x(iT)-a_{\xi}.$$

Приведенная оценка называется периодограммой. Прежде всего приведем эту оценку к более понятному виду. Для этого возведем $A(\omega)$ и $B(\omega)$ в квадрат и подставим в выражение (8.8). Получим

$$\widehat{s}_{\xi,T}(\omega)=\frac{T}{2\pi n}(\sum_{i=1}^{n}\sum_{j=1}^{n}\overset{\circ}{x}(iT)\overset{\circ}{x}(jT)\cos(i\omega T)\cos(j\omega T))+$$

$$+\sum_{i=1}^{n}\sum_{j=1}^{n}\overset{\circ}{x}(iT)\overset{\circ}{x}(jT)\sin(i\omega T)\sin(j\omega T))=$$

$$=\frac{T}{2\pi n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\overset{\circ}{x}(iT)\overset{\circ}{x}(jT)\big(\cos(i\omega T)\cos(j\omega T)+\sin(i\omega T)\sin(j\omega T)\big)\right)=$$

$$=\frac{T}{2\pi n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\overset{\circ}{x}(iT)\overset{\circ}{x}(jT)\cos((i-j)\omega T)\right).$$

Последняя сумма есть сума элементов квадратной $(n \times n)$-матрицы, причем суммирование ведется по строкам. Будем выполнять суммирование этих элементов иначе. Введем новую переменную суммирования $k = j - i$. Понятно, что k изменяется от $-(n-1)$ до $(n-1)$. Переменная k упорядочивает диагонали нашей матрицы от левой нижней до правой верхней. Число элементов в k-й диагонали равно $n-k$. Суммируя элементы нашей матрицы по диагоналям, получим

$$\hat{s}_{\xi,T}(\omega) = \frac{T}{2\pi} \sum_{k=-(n-1)}^{n-1} \left(\frac{1}{n} \sum_{i=1}^{n-k} \overset{\circ}{x}(iT) \overset{\circ}{x}(iT + kT) \right) \cos(k\omega T) =$$

$$= \frac{T}{2\pi} \sum_{k=-(n-1)}^{n-1} \hat{R}_{\xi}^{*}(kT) \cos(k\omega T) = \frac{T}{2\pi} \sum_{k=-(n-1)}^{n-1} (1 - \frac{|k|}{n}) \hat{R}_{\xi}(kT) \cos(k\omega T). \quad (8.9)$$

Последнее выражение позволяет, во-первых, увидеть аналогию между оценкой и оцениваемой величиной, которая имеет вид

$$s_{\xi,T}(\omega) = \frac{T}{2\pi} \sum_{k=-\infty}^{\infty} R_{\xi}(kT) \cos(k\omega T).$$

Во-вторых, мы легко найдем математическое ожидание оценки:

$$E(\hat{s}_{\xi,T}(\omega)) = \frac{T}{2\pi} \sum_{k=-(n-1)}^{n-1} (1 - \frac{|k|}{n}) R_{\xi}(kT) \cos(k\omega T).$$

Учитывая формулу для коэффициента ковариации

$$R_{\xi}(kT) = \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} s_{\xi,T}(v) \cos(kvT) dv,$$

мы получим

$$E(\hat{s}_{\xi,T}(\omega)) = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \sum_{k=-(n-1)}^{n-1} (1 - \frac{|k|}{n}) s_{\xi,T}(v) \cos(kvT) \cos(k\omega T) dv.$$

Видно, что математическое ожидание оценки $E(\hat{s}_{\xi,T}(\omega))$ не равно оцениваемой функции $s_{\xi,T}(\omega)$, а является взвешенным интегралом от этой функции. Следовательно, оценка $\hat{s}_{\xi,T}(\omega)$ смещенная. Можно, однако, доказать ее асимптотическую несмещенность. В частности, справедлива следующая теорема [1].

Теорема 8.2. Если $\sum_{k=-\infty}^{\infty} R_{\xi}(kT)\cos(k\omega T) < \infty$, то

$$\lim_{n\to\infty} E\left(\hat{s}_{\xi,T}(\omega)\right) = -\frac{T}{2\pi}\sum_{k=-\infty}^{\infty} R_{\xi}(kT)\cos(k\omega T).$$

Если $s_{\xi,T}(\omega)$ непрерывна в рассматриваемой точке ω, то

$$\lim_{n\to\infty} E\left(\hat{s}_{\xi,T}(\omega)\right) = s_{\xi,T}(\omega).$$

Из данной теоремы следует, что в случае непрерывности спектральной плотности $s_{\xi,T}(\omega)$ ее оценка $\hat{s}_{\xi,T}(\omega)$ является асимптотически несмещенной.

Для дисперсии оценки $\hat{s}_{\xi,T}(\omega)$ справедлива следующая теорема [1].

Теорема 8.3. Если

$$\lim_{n\to\infty} \frac{1}{n} \sum_{r,u,w=-(n-1)}^{n-1} |\kappa(r,u,w)| = 0,$$

где $\kappa(r,u,w)$ – семиинвариант четвертого порядка стационарной случайной последовательности $\xi(kT)$, и $s_{\xi,T}(\omega)$ непрерывна при $\omega = 0$ и $\omega = \pi/T$, то

$$\lim_{n\to\infty} D(\hat{s}_{\xi,T}(0)) = 2s^2_{\xi,T}(0), \ \lim_{n\to\infty} D(\hat{s}_{\xi,T}(\pm\pi/T)) = 2s^2_{\xi,T}(\pi/T).$$

Если $s_{\xi,T}(\omega)$ непрерывна при $\omega = \lambda$, то

$$\lim_{n\to\infty} D(\hat{s}_{\xi,T}(\lambda)) = 2s^2_{\xi,T}(\lambda), \ \lambda \neq 0, \ \lambda \neq \pm\pi/T.$$

Важно заметить, что, в соответствии с теоремой, дисперсия оценки $\hat{s}_{\xi,T}(\omega)$ не стремиться к нулю при $n \to \infty$. Это свидетельствует о том, что оценка $\hat{s}_{\xi,T}(\omega)$ не является состоятельной для $s_{\xi,T}(\omega)$.

8.4.2 Случай неизвестного математического ожидания

Рассмотрим теперь случай, когда математическое ожидание неизвестно. В этом случае можно использовать оценку того же вида, что и (8.8)

$$\bar{s}_{\xi,T}(\omega)=\frac{nT}{8\pi}(A^2(\omega)+B^2(\omega)),\quad -\frac{\pi}{T}\le\omega\le\frac{\pi}{T}, \tag{8.10}$$

где

$$A(\omega)=\frac{2}{n}\sum_{i=1}^{n}\overset{*}{x}(iT)\cos(i\omega T),$$

$$B(\omega)=\frac{2}{n}\sum_{i=1}^{n}\overset{*}{x}(iT)\sin(i\omega T),$$

$$\overset{*}{x}(iT)=x(iT)-\hat{a},$$

и $\hat{a}$ – оценка математического ожидания (8.2).

Прежде всего отметим, что $\bar{s}_{\xi,T}(0)=0$ для любой реализации $x(T),x(2T),...,x(nT)$. Действительно, при $\omega=0$ $\sin(i\omega T)=0$, поэтому $B(\omega)=0$. Покажем, что и $A(0)=0$. Преобразуя $A(\omega)$, получим

$$A(\omega)=\frac{2}{n}\sum_{i=1}^{n}(x(iT)-\hat{a})=\frac{2}{n}\sum_{i=1}^{n}\left(x(iT)-\frac{1}{n}\sum_{j=1}^{n}x(jT)\right)=$$

$$=2\left(\frac{1}{n}\sum_{i=1}^{n}x(iT)-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x(jT)\right)=2\left(\frac{1}{n}\sum_{i=1}^{n}x(iT)-\frac{1}{n}\sum_{j=1}^{n}x(jT)\right)=0.$$

Это значит, что $\bar{s}_{\xi,T}(\omega)$ при $\omega=0$ вообще не является оценкой, так как всегда равна нулю. Тем не менее, при $\omega\neq 0$ оценка $\bar{s}_{\xi,T}(\omega)$ асимптотически несмещенная. Это вытекает из следующей теоремы [1].

Теорема 8.4. Если $s_{\xi,T}(\omega)$ непрерывна при $\omega=0$ и $\omega=\lambda$, то

$$\lim_{n\to\infty} n\left(E(\bar{s}_{\xi,T}(\lambda))-E(\hat{s}_{\xi,T}(\lambda))\right)=0,\quad \lambda\neq 0.$$

Для дисперсии оценки $\bar{s}_{\xi,T}(\omega)$ существует следующее асимптотическое соотношение: если семиинвариант четвертого порядка $\kappa(r,u,w)=0$, то

$$\lim_{n\to\infty} n\left(D(\bar{s}_{\xi,T}(\lambda))-D(\hat{s}_{\xi,T}(\lambda))\right)=0,\ \lambda\neq 0.$$

Данное соотношение свидетельствует о том, что оценка $\bar{s}_{\xi,T}(\omega)$ не лучше оценки $\hat{s}_{\xi,T}(\omega)$, то есть также не состоятельна.

8.5 Оценки спектральной плотности, основанные на оценках ковариаций

Рассмотренные нами ранее оценки являются достаточно плохими. Поэтому применяются другие оценки. В следующих разделах рассматриваются обобщенные оценки спектральной плотности.

8.5.1 Случай известного математического ожидания

Для случая известного математического ожидания обобщенную оценку запишем их в виде

$$\hat{\bar{s}}_{\xi,T}(\omega)=\frac{T}{2\pi}\sum_{k=-(n-1)}^{n-1} w_k^*\hat{R}_\xi^*(kT)\cos(k\omega T)=\frac{T}{2\pi}\sum_{k=-(n-1)}^{n-1} w_k\hat{R}_\xi(kT)\cos(k\omega T) \quad (8.11)$$

где w_k^* – надлежащим образом выбранные числа, зависящие от n, а числа w_k определяются из (8.7) выражением

$$w_k=\left(1-\frac{|k|}{n}\right)w_k^*.$$

Оценка $\hat{s}_{\xi,T}(\omega)$ (8.9) является частным случаем оценки $\hat{\bar{s}}_{\xi,T}(\omega)$ (8.11), когда $w_k^*=1$. Если обозначить

$$\widehat{R}^{*}_{\xi}(kT)=\int\limits_{-\frac{\pi}{T}}^{\frac{\pi}{T}}\widehat{s}_{\xi,T}(v)\cos(kvT)dv,$$

то оценку $\widehat{\widehat{s}}_{\xi,T}(\omega)$ (8.11) можно записать в виде

$$\widehat{\widehat{s}}_{\xi,T}(\omega)=\frac{T}{2\pi}\int\limits_{-\frac{\pi}{T}}^{\frac{\pi}{T}}\left(\sum_{k=-(n-1)}^{n-1}w^{*}_{k}\cos(k\omega T)\cos(kvT)\right)\widehat{s}_{\xi,T}(v)dv,$$

или в виде

$$\widehat{\widehat{s}}_{\xi,T}(\omega)=\int\limits_{-\frac{\pi}{T}}^{\frac{\pi}{T}}w^{*}(v,\omega)\widehat{s}_{\xi,T}(v)dv, \qquad (8.12)$$

где

$$w^{*}(v,\omega)=\frac{T}{2\pi}\left(\sum_{k=-(n-1)}^{n-1}w^{*}_{k}\cos(k\omega T)\cos(kvT)\right). \qquad (8.13)$$

Функция $w^{*}(v,\omega)$ (8.13) называется частотным окном, а оценка $\widehat{\widehat{s}}_{\xi,T}(\omega)$ (8.12) представляет собой взвешенное среднее оценки $\widehat{s}_{\xi,T}(\omega)$ (периодограммы).

8.5.2 Случай неизвестного математического ожидания

Для случая неизвестного математического ожидания обобщенную оценку запишем в виде

$$\overline{\widehat{s}}_{\xi.T}(\omega)=\frac{T}{2\pi}\sum_{k=-(n-1)}^{n-1}w^{*}_{k}\overline{R}^{*}_{\xi}(kT)\cos(k\omega T)=\frac{T}{2\pi}\sum_{k=-(n-1)}^{n-1}w_{k}\overline{R}_{\xi}(kT)\cos(k\omega T). \qquad (8.14)$$

Если обозначить

$$\overline{R}^{*}_{\xi}(kT) = \int\limits_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \overline{s}_{\xi,T}(\nu)\cos(k\nu T)d\nu,$$

то эту оценку можно представить как

$$\overline{\overline{s}}_{\xi,T}(\omega) = \frac{T}{2\pi}\int\limits_{-\frac{\pi}{T}}^{\frac{\pi}{T}}\left(\sum_{k=-(n-1)}^{n-1} w^{*}_{k}\cos(k\omega T)\cos(k\nu T)\right)\overline{s}_{\xi,T}(\nu)d\nu$$

или

$$\overline{\overline{s}}_{\xi,T}(\omega) = \int\limits_{-\frac{\pi}{T}}^{\frac{\pi}{T}} w^{*}(\nu,\omega)\,\overline{s}_{\xi,T}(\nu)d\nu, \qquad (8.15)$$

где $w^{*}(\nu,\omega)$ определяется выражением (8.13), а $\overline{s}_{\xi,T}(\nu)$ – выражением (8.10).

8.5.3 Примеры оценок спектральной плотности

Выбирая в формулах (8.11), (8.14) различные веса w_k, w^{*}_{k}, или, иначе, в формулах (8.12), (8.15) различные окна $w^{*}(\nu,\omega)$, можно получить различные оценки спектральной плотности. Приведем некоторые оценки.

1. Периодограмма. Если выбрать

$$w^{*}_{k} = 1,\ w_k = \left(1 - \frac{|k|}{n}\right),\ k = 0,\pm1,...,\pm(n-1),$$

то мы получим рассмотренную нами ранее периодограмму.

2. Усеченная оценка.. Если выбрать некоторое число $m \le (n-1)$ и назначить веса следующим образом

$$w^{*}_{k} = \begin{cases} 1 & для\ \ k = 0,\pm1,...,\pm m, \\ 0 & для\ \ k = \pm(m+1),...,\pm(n-1), \end{cases}$$

или, что то же,

$$w_k = \begin{cases} 1 - \dfrac{|k|}{n} & для \;\; k = 0, \pm 1, ..., \pm m, \\ 0 & для \;\; k = \pm(m+1), ..., \pm(n-1), \end{cases}$$

то мы получим так называемую усеченную оценку. Выражение этой оценки имеет вид

$$\hat{\bar{s}}_{\xi,T}(\omega) = \frac{T}{2\pi} \sum_{k=-m}^{m} \hat{R}^*_{\xi}(kT) \cos(k\omega T) = \frac{T}{2\pi} \sum_{k=-m}^{m} (1 - \frac{|k|}{n}) \hat{R}_{\xi}(kT) \cos(k\omega T).$$

3. Оценка Бартлетта. Выбирая число $m \le (n-1)$ и веса вида

$$w^*_k = \begin{cases} 1 - \dfrac{|k|}{m} & для \;\; k = 0, \pm 1, ..., \pm m, \\ \dfrac{1 - \dfrac{|k|}{m}}{1 - \dfrac{|k|}{n}} & \\ 0 & для \;\; k = \pm(m+1), ..., \pm(n-1), \end{cases}$$

или, что то же самое,

$$w_k = \begin{cases} 1 - \dfrac{|k|}{m} & для \;\; k = 0, \pm 1, ..., \pm m, \\ 0 & для \;\; k = \pm(m+1), ..., \pm(n-1), \end{cases}$$

мы получим так называемую оценку Бартлетта. Выражение этой оценки имеет вид

$$\hat{\bar{s}}_{\xi,T}(\omega) = \frac{T}{2\pi} \sum_{k=-m}^{m} (1 - \frac{|k|}{m}) \hat{R}_{\xi}(kT) \cos(k\omega T).$$

3. Модифицированная оценка Бартлетта. Если в предыдущую оценку ввести еще один сглаживающий множитель $1 - |k| / n$, то в результате получим оценку

$$\hat{\bar{s}}_{\xi,T}(\omega) = \frac{T}{2\pi} \sum_{k=-m}^{m} (1 - \frac{|k|}{m})(1 - \frac{|k|}{n}) \hat{R}_{\xi}(kT) \cos(k\omega T).$$

Коэффициенты в этом случае равны

$$w_k = \begin{cases} \left(1 - \dfrac{|k|}{m}\right)\left(1 - \dfrac{|k|}{n}\right) & для \;\; k = 0, \pm 1, ..., \pm m, \\ 0 & для \;\; k = \pm(m+1), ..., \pm(n-1), \end{cases}$$

$$w_k^* = \begin{cases} \left(1 - \dfrac{|k|}{m}\right) & \text{для } k = 0, \pm 1, ..., \pm m, \\ 0 & \text{для } k = \pm(m+1), ..., \pm(n-1). \end{cases}$$

Существуют также оценки Даниэля, Хеннинга, Хемминга, Парзена и др. [1].

Принцип выбора весовых коэффициентов состоит в следующем. Поскольку оценки коэффициентов ковариации $\widehat{R}_\xi(kT)$ при больших значениях k определяются по небольшому числу $(n-k)$ отсчетов и поэтому являются плохими, то их влияние на оценку спектральной плотности желательно ограничить, учитывая их с небольшими или вообще с нулевыми весовыми коэффициентами.

9 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ДЛЯ ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН

В корреляционном анализе изучается взаимосвязь между случайными величинами на основе экспериментальных (эмпирических) данных. Случайные величины ξ, η считаются распределенными по совместному нормальному закону. Требуется по выборке $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$ из этого двухмерного распределения сделать статистические выводы о случайных величинах ξ, η и их взаимосвязи. Теоретической основой корреляционного анализа является следующая вероятностная теорема.

Теорема 9.1. Если случайные величины ξ, η имеют совместное нормальное распределение с математическими ожиданиями a_ξ, a_η, дисперсиями σ_ξ^2, σ_η^2 и коэффициентом корреляции $r_{\xi\eta}$, то условная плотность вероятности $f_{\eta/\xi}(y/x)$ также нормальная следующего вида:

$$f_{\eta/\xi}(y/x) = N\left(a_\eta + r_{\xi\eta}\frac{\sigma_\eta}{\sigma_\xi}(x - a_\xi), \sigma_\eta^2(1 - r_{\xi\eta}^2)\right),$$

где первый параметр этого распределения – условное математическое ожидание или функции регрессии

$$a_{\eta/\xi}(x) = a_\eta + r_{\xi\eta}\frac{\sigma_\eta}{\sigma_\xi}(x - a_\xi),$$

а второй – условная дисперсия

$$\sigma^2 = \sigma_{\eta/\xi}^2(x) = \sigma_\eta^2(1 - r_{\xi\eta}^2).$$

В корреляционном анализе решаются следующие задачи: 1) отыскание оценок параметров функции регрессии; 2) отыскание распределений этих оценок; 3) проверка гипотез о значимости параметров функции регрессии и построение для них доверительных интервалов. Приведем решение этих задач.

9.1 Оценки параметров функции регрессии

В соответствии с приведенной выше теоремой теоретическая функция регрессии имеет вид

$$a_{\eta/\xi}(x) = a_\eta + b(x - a_\xi),\ b = r_{\xi\eta}\frac{\sigma_\eta}{\sigma_\xi}. \tag{9.1}$$

Один из путей получения эмпирической функции регрессии состоит в замене теоретических характеристик выборочными (эмпирическими) характеристиками (выборочный метод). В результате такой замены получим эмпирическую функцию регрессии

$$\hat{a}_{\eta/\xi}(x) = \hat{a}_\eta + \hat{b}(x - \hat{a}_\xi),$$

которая является оценкой теоретической функции регрессии (9.1). Оценки параметров функции регрессии определяются выражениями (см. раздел 1.8):

$$\hat{a}_\eta = \bar{y},\ \hat{b} = \bar{r}_{xy}\frac{\bar{s}_y}{\bar{s}_x} = \frac{\bar{s}_{xy}}{\bar{s}_x^2},\ \hat{a}_\xi = \bar{x}, \tag{9.2}$$

$$\bar{s}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{xy}, \tag{9.3}$$

$$\bar{s}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \tag{9.4}$$

$$\bar{s}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}y_i^2 - \bar{y}^2,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i, \tag{9.5}$$

$$\hat{\sigma}^2 = \bar{s}_y^2(1 - \bar{r}_{xy}^2), \tag{9.6}$$

$$\bar{r}_{xy} = \frac{\bar{s}_{xy}}{\bar{s}_x\bar{s}_y}. \tag{9.7}$$

9.2 Распределения оценок параметров функции регрессии

Относительно распределений оценок и их свойств в литературе доказаны следующие утверждения (см., например, [12]).

Случайные векторы $(\bar{x}, \bar{y})$ и $(\bar{s}_x^2, \bar{s}_{xy}, \bar{s}_y^2)$ независимы. Оценка $\hat{a}_\eta \in N(a_\eta, \frac{\sigma_\eta^2}{n})$, является несмещенной, состоятельной, эффективной.

Аналогично оценка $\hat{a}_\xi \in N(a_\xi, \frac{\sigma_\xi^2}{n})$, несмещенная, состоятельная, эффективная. Приведем распределения других статистик:

$$u_\eta = \frac{\hat{a}_\eta - a_\eta}{\sigma_\eta}\sqrt{n} \in N(0,1),$$

$$u_\xi = \frac{\hat{a}_\xi - a_\xi}{\sigma_\xi}\sqrt{n} \in N(0,1),$$

$$v = \frac{n\bar{s}_y^2}{\sigma_\eta^2} \in H_1(n-1),$$

$$t_{a_\eta} = \frac{\hat{a} - a_\eta}{\bar{s}_y}\sqrt{n-1} \in T_1(n-1),$$

$$w = \frac{n\bar{s}_x^2}{\sigma_\xi^2} \in H_1(n-1),$$

$$t_{a_\xi} = \frac{\hat{a}_\xi - a_\xi}{\bar{s}_x}\sqrt{n-1} \in T_1(n-1).$$

Если теоретический коэффициент корреляции $r_{\xi\eta} = 0$, то

$$t_r = \frac{\bar{r}_{xy}}{\sqrt{1-\bar{r}_{xy}^2}}\sqrt{n-2} = \frac{\bar{r}_{xy}}{\hat{\sigma}}\bar{s}_y\sqrt{n-2} \in T_1(n-2).$$

Оценка $\bar{r}_{xy}$ коэффициента корреляции $r_{\xi\eta}$ асимптотически несмещенная и состоятельная. Статистика

$$t_b = \frac{\hat{b} - b}{\hat{\sigma}} \bar{s}_x \sqrt{n-2} \in T_1(n-2).$$

9.3 Проверка гипотез и построение доверительных интервалов для параметров функции регрессии

Приведенные распределения позволяют выполнять проверку гипотез относительно соответствующих параметров функции регрессии и построение доверительных интервалов для них. В частности, гипотеза вида $\{H_0 : r_{\xi\eta} = 0;$ $H_1 : r_{\xi\eta} \neq 0\}$ – это гипотеза о значимости коэффициента корреляции между случайными величинами ξ, η. Она проверяется на основе статистики t_r при двухстороннем критерии значимости

$$P(|t_r| > t_{\frac{\alpha}{2}}) = \alpha.$$

Гипотеза вида $\{H_0 : a_\eta = a_0;\ H_1 : a_\eta \neq a_0\}$ проверяется на основе статистики t_{a_η}. Гипотеза $\{H_0 : b = b_0;\ H_1 : b \neq b_0\}$ проверяется на основе статистики t_b. Гипотезы вида $\{H_0 : a_\eta = 0;\ H_1 : a_\eta \neq 0\}$ или $\{H_0 : b = 0;\ H_1 : b \neq 0\}$ называются гипотезами о значимости коэффициентов регрессии. Если при проверке такой гипотезы коэффициент признан незначимым, то в зависимости (9.1) им можно пренебречь.

10 РЕГРЕССИОННЫЙ АНАЛИЗ

В регрессионном анализе изучается взаимосвязь между случайными и неслучайными величинами на основе экспериментальных данных. В отличие от корреляционного анализа в регрессионном анализе не все изучаемые взаимосвязанные величины являются случайными, и условия, накладываемые на изучаемые величины, менее обременительны. В связи с этим считается, что задачи регрессионного анализа чаще встречаются на практике по сравнению с задачами корреляционного анализа.

10.1 Линейный регрессионный анализ для двух переменных (простая линейная регрессия)

Основное предположение регрессионного анализа состоит в следующем.

Случайная величина η зависит от неслучайной (детерминированной) величины x таким образом, что условная плотность вероятности $f_\eta(y/x)$ нормальная и имеет вид

$$f_\eta(y/x) = N(\varphi(x), \sigma^2)),$$

где $\varphi(x)$ – условное математическое ожидание или функция регрессии величины η на x, σ^2 – условная дисперсия, не зависящая от x. Функция регрессии $\varphi(x)$ обычно задается с точностью до некоторых параметров $\theta_1,...,\theta_m$ в виде $\varphi(x,\theta_1,...,\theta_m)$. Имеются результаты наблюдений над величинами η и x в виде совокупности пар чисел

$$(x_1, y_1),(x_2, y_2),...,(x_n, y_n),$$

где $y_1,...,y_n$ – значения величины η, соответствующие значениям $x_1,...,x_n$ величины x. Требуется по этим данным найти оценки $\hat{\theta}_1,...,\hat{\theta}_m$ параметров $\theta_1,...,\theta_m$ функции регрессии $\varphi(x,\theta_1,...,\theta_m)$ и параметра σ^2 распределения

$f_\eta(y/x) = N(\varphi(x), \sigma^2)$, а также сделать статистические выводы о этих параметрах.

Рассматриваемую в регрессионном анализе модель зависимости между η и x можно представить в виде

$$\eta = \varphi(x, \theta_1, ..., \theta_m) + \xi,$$

где ξ – случайная величина, имеющая распределение $N(0, \sigma^2)$. Придавая переменной x значения $x_1, ..., x_n$ из некоторого промежутка, мы получим значения $y_1, ..., y_n$ случайной величины η:

$$y_i = \varphi(x, \theta_1, ..., \theta_m) + z_i,$$

где z_i – возможные значения случайной величины ξ (ошибки измерений).

Рассмотрим случай линейной функции регрессии

$$\varphi(x) = \alpha + \beta x$$

с двумя параметрами α и β. Для удобства данную функцию регрессии будем рассматривать в виде

$$\varphi(x) = a + b(x - \bar{x}),$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – среднее арифметическое наблюдений переменной x. Понятно, что в этом случае $\alpha = a - b\bar{x}$, $\beta = b$.

Для получения оценок параметров a, b функции регрессии по результатам экспериментов, то есть для получения эмпирической функции регрессии, воспользуемся методом наименьших квадратов. В этом случае оценки $\hat{a}$, $\hat{b}$ параметров a, b определяются как решение следующей оптимизационной задачи:

$$F(a,b) = \sum_{i=1}^{n} \left(y_i - a - b(x_i - \bar{x})\right)^2 \to \min_{a,b}.$$

Необходимые условия минимума функции $F(a,b)$ имеют вид системы уравнений

$$\frac{\partial}{\partial a}F(a,b)=0,\ \frac{\partial}{\partial b}F(a,b)=0.$$

В результате дифференцирования функции $F(a,b)$ получим следующую систему уравнений

$$-2\sum_{i=1}^{n}\left(y_i-a-b(x_i-\bar{x})\right)=0,$$

$$-2\sum_{i=1}^{n}\left(y_i-a-b(x_i-\bar{x})\right)(x_i-\bar{x})=0.$$

Упростим эту систему:

$$na+nb\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}y_i\ ,$$

$$a\sum_{i=1}^{n}(x_i-\bar{x})+b\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}y_i(x_i-\bar{x}).$$

Поскольку $\sum_{i=1}^{n}(x_i-\bar{x})=0$, то получим

$$na=\sum_{i=1}^{n}y_i\ ,$$

$$b\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}y_ix_i-\bar{x}\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{n}y_ix_i-\overline{xy}.$$

Отсюда мы получаем следующие оценки для неизвестных параметров:

$$\hat{a}=\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i\ ,$$

$$\hat{b}=\bar{r}_{xy}\frac{\bar{s}_y}{\bar{s}_x}=\frac{\bar{s}_{xy}}{\bar{s}_x^2},$$

$$\hat{\sigma}^2=\bar{s}_y^2(1-\bar{r}_{xy}^2).$$

Выражения для оценок $\bar{r}_{xy}$, $\bar{s}_x$, $\bar{s}_y$ такие же, как и в разделе 9.1. Метод наименьших квадратов не позволяет непосредственно получить оценку дисперсии

σ^2 ошибок измерений. Эта оценка определяется из дополнительных соображений. Естественно в качестве такой оценки взять величину

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}(x_i - \bar{x}))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 ,$$

где обозначено

$$\hat{y}_i = \hat{a} - \hat{b}(x_i - \bar{x}).$$

Эта оценка совпадает с оценкой (9.6). Действительно, преобразуя выражение для $\hat{\sigma}^2$ так, как показано ниже,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}(x_i - \bar{x}))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y} - \frac{\bar{s}_{xy}}{\bar{s}_x^2}(x_i - \bar{x}))^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\overset{\circ}{y}_i - \frac{\bar{s}_{xy}}{\bar{s}_x^2}\overset{\circ}{x}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\overset{\circ}{y}_i)^2 - \frac{2}{n}\frac{\bar{s}_{xy}}{\bar{s}_x^2}\sum_{i=1}^{n}\overset{\circ}{x}_i\overset{\circ}{y}_i + \frac{1}{n}\frac{\bar{s}_{xy}^2}{\bar{s}_x^4}\sum_{i=1}^{n}(\overset{\circ}{x}_i)^2 =$$

$$= \bar{s}_y^2 - \frac{\bar{s}_{xy}^2}{\bar{s}_x^2} = \bar{s}_y^2\left(1 - \frac{\bar{s}_{xy}^2}{\bar{s}_x^2\bar{s}_y^2}\right) = \bar{s}_y^2(1 - \bar{r}_{xy}^2),$$

мы получаем выражение (9.6). Таким образом, мы получили те же оценки (9.2) – (9.7), что и в корреляционном анализе. Рассмотрим свойства этих оценок.

Оценки $\hat{a}$, $\hat{b}$ – несмещенные, $\hat{\sigma}^2$ – смещенная, однако асимптотически несмещенная. Новая оценка

$$\hat{\sigma}_1^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}(x_i - \bar{x}))^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

является несмещенной. Легко видеть, что $\hat{\sigma}_1^2 = \frac{n}{n-2}\hat{\sigma}^2$.

Рассмотрим распределения оценок. Можно показать, что

$$\hat{a} \in N(a, \frac{\sigma^2}{n}),$$

$$\hat{b} \in N(b, \frac{\sigma^2}{\bar{s}_x^2 n}),$$

$$u_1 = \frac{\hat{a} - a}{\sigma}\sqrt{n} \in N(0,1),$$

$$u_2 = \frac{\hat{b} - b}{\sigma}\bar{s}_x\sqrt{n} \in N(0,1),$$

$$v = \frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-2)\hat{\sigma}_1^2}{\sigma^2} \in H_1(n-2).$$

Оценки $\hat{a}$, $\hat{b}$, $\hat{\sigma}^2$ независимы. Отсюда можно сделать вывод, что

$$t_a = \frac{\hat{a} - a}{\hat{\sigma}}\sqrt{n-2} \in T_1(n-2),$$

$$t_b = \frac{\hat{b} - b}{\hat{\sigma}}\bar{s}_x\sqrt{n-2} = \frac{\hat{b} - b}{\hat{\sigma}_1}\bar{s}_x\sqrt{n} \in T_1(n-2).$$

Статистики t_a, t_b, v применяются для построения доверительных интервалов для параметров a, b, σ^2 соответственно и для проверки гипотез относительно этих параметров, например гипотезы $\{H_0 : a = a_0;\ H_1 : a \neq a_0\}$.

10.2 Регрессионный анализ для многих переменных (множественная нелинейная регрессия)

10.2.1 Постановка задачи

Рассмотрим некоторый исследуемый объект в виде черного ящика с векторным входом и скалярным выходом. Будем считать, что выходная переменная объекта η зависит от входной векторной переменной $X = (x_1, x_2, ..., x_q)$ таким образом, что условная плотность вероятности величины η нормальная,

$$f(\eta / X) = N(\varphi(X), \sigma^2),$$

где

$$y = \varphi(X)$$

– условное математическое ожидание (функция регрессии η на X), σ^2 – условная дисперсия. Эту зависимость можно представить в виде

$$\eta = \varphi(X) + \xi,$$

где случайная величина ξ имеет нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2: $\xi \in N(0, \sigma^2)$.

Пусть для некоторых значений $X_1,...,X_n$ входной векторной переменной X получены значения $y_{o,1},...,y_{o,n}$ выходной случайной переменной η,

$$y_{o,i} = \varphi(X_i) + z_i,\ i = \overline{1,n},$$

где z_i – значения случайной величины ξ. Требуется по наблюдениям $(X_1, y_{o1}),...,(X_n, y_{on})$ построить математическую модель объекта, то есть получить эмпирическую функцию регрессии $\hat{y} = \hat{\varphi}(X)$ и оценку $\hat{\sigma}^2$ условной дисперсии σ^2. Понятно, что построенная модель будет отличаться от действительной, однако это отличие должно быть несущественным (незначимым).

Обычно для аппроксимации (подбора) функции регрессии используют класс функций, линейных по параметрам, то есть представимых в виде

$$y = \varphi(X, \bar{\theta}) = \sum_{j=1}^{m} \theta_j h_j(X), \qquad (10.1)$$

где $h_j(X)$ – некоторые функции вектора X, а θ_j, $j = \overline{1,m}$, – неизвестные параметры, $\bar{\theta} = (\theta_1,...,\theta_m)$ – вектор неизвестных параметров функции регрессии $\varphi(X, \bar{\theta})$. Функция (10.1) называется *гипотетической* функцией регрессии. (Функцию регрессии, полученную по результатам экспериментов, называют *эмпирической)*.

Гипотетическую функцию регрессии (10.1) можно записать в векторно-матричной форме

$$y = \varphi(X, \bar{\theta}) = H^T \bar{\theta} = \bar{\theta}^T H, \qquad (10.2)$$

где $H^T = (h_j(X))$, $j = \overline{1,m}$, – вектор-строка функций от X, $\bar{\theta}^T = (\theta_j)$, $j = \overline{1,m}$, – вектор-строка параметров, m – число неизвестных параметров. Отметим, что функция, линейная по параметрам, может быть как линейной, так и нелинейной по своим аргументам $X = (x_1, x_2, ..., x_q)$. Например, желая получить линейную по входной скалярной переменной x функцию регрессии, мы должны выбрать

$$H^T = (1, x),\ \bar{\theta}^T = (\alpha, \beta).$$

Тогда

$$y = \varphi(x, \alpha, \beta) = (1, x)\begin{pmatrix}\alpha \\ \beta\end{pmatrix} = \alpha + \beta x.$$

Желая получить функцию регрессии скалярной переменной x в виде полинома второй степени $y = \alpha + \beta x_1 + \gamma x_2 + \tau x_1^2$, необходимо выбрать

$$H^T = (1, x_1, x_2, x_1^2),\ \bar{\theta}^T = (\alpha, \beta, \gamma, \tau).$$

Из приведенных примеров легко заметить, что в любой линейной по параметрам модели, содержащей свободный член θ_1, функция $h_1(X)$ должна быть равной единице.

Часто выбирают модель, линейную как по параметрам, так и по входным переменным. В этом случае

$$H^T = H^T(X) = X^T = (x_1, x_2, ..., x_m),$$

и функция регрессии имеет вид:

$$y = \varphi(X, \bar{\theta}) = \sum_{j=1}^{m} \theta_j x_j = X^T \bar{\theta} = \bar{\theta}^T X.$$

Если такая модель должна содержать свободный член, то необходимо выбрать $x_1 = 1$.

10.2.2 Оценки параметров функции регрессии

В соответствии с гипотетической моделью (10.1) (или (10.2)) мы имеем следующий набор данных

$$y_{o,i} = H_i^T \bar{\theta} + z_i, \quad i = \overline{1,n}, \tag{10.3}$$

где $H_i^T = (h_{i,j}) = (h_j(X_i)), \ i = \overline{1,n}, \ z_i$ – независимые случайные величины с нормальным распределением $N(0,\sigma^2)$. В этих условиях случайные величины y_{oi} распределены по нормальному закону $N(H_i^T \bar{\theta}, \sigma^2) = N(y_i, \sigma^2)$, то есть имеют плотность вероятности вида

$$f(y_{o,i}, \bar{\theta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_{o,i} - H_i^T \bar{\theta})^2 \right), \ i = \overline{1,n}. \tag{10.4}$$

Для нахождения оценок воспользуемся методом максимума правдоподобия. Функция правдоподобия для выборки (10.3) при распределении (10.4) имеет вид:

$$L(\bar{\theta}, \sigma^2) = \prod_{i=1}^{n} f(y_{o,i}, \bar{\theta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_{o,i} - H_i^T \bar{\theta})^2 \right) =$$

$$= \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_{o,i} - H_i^T \bar{\theta})^2 \right),$$

а логарифмическая функция правдоподобия – вид:

$$ln\, L(\bar{\theta}, \sigma^2) = -ln\sqrt{(2\pi)^n} - n\, ln\, \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_{o,i} - H_i^T \bar{\theta})^2 .$$

Необходимые условия максимума функции $\ln L(\bar{\theta}, \sigma^2)$ состоят в следующем:

$$\frac{\partial}{\partial \bar{\theta}} ln\, L(\bar{\theta}, \sigma^2) = 0,$$

$$\frac{\partial}{\partial \sigma^2} ln\, L(\bar{\theta}, \sigma^2) = 0.$$

Перепишем выражение для $ln\,L(\bar{\theta},\sigma^2)$, отбросив слагаемое, не зависящее от параметров:

$$ln\,L(\bar{\theta},\sigma^2) \sim -n\,ln\,\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{o,i} - H_i^T\bar{\theta})^T(y_{o,i} - H_i^T\bar{\theta}) =$$

$$= -n\,ln\,\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{o,i}^2 - 2y_{o,i}H_i^T\bar{\theta} + \bar{\theta}^T H_i H_i^T\bar{\theta}).$$

Дифференцирование скалярной функции $ln\,L(\bar{\theta},\sigma^2)$ по параметрам $\bar{\theta}$ и σ^2 дает нам следующие уравнения

$$\sum_{i=1}^{n}(H_i y_{o,i} - H_i H_i^T\bar{\theta}) = 0, \qquad (10.5)$$

$$n\sigma^2 - \sum_{i=1}^{n}(y_{o,i} - H_i^T\bar{\theta})^2 = 0, \qquad (10.6)$$

которые называются нормальными уравнениями. Первое из этих уравнений можно переписать в виде

$$(\sum_{i=1}^{n} H_i H_i^T)\bar{\theta} = \sum_{i=1}^{n} H_i y_{o,i}\,.$$

Будем считать векторы $H_1,...,H_n$ линейно независимыми, что обеспечивает невырожденность матрицы $\sum_{i=1}^{n} H_i H_i^T$ и существование обратной матрицы $(\sum_{i=1}^{n} H_i H_i^T)^{-1}$. Умножим обе части последнего уравнения слева на матрицу $(\sum_{i=1}^{n} H_i H_i^T)^{-1}$, в результате чего получим следующую оценку вектора-столбца параметров $\bar{\theta}$:

$$\hat{\bar{\theta}} = (\sum_{i=1}^{n} H_i H_i^T)^{-1}(\sum_{i=1}^{n} H_i y_{o,i}). \qquad (10.7)$$

С учетом этой оценки из второго нормального уравнения (10.6) получаем оценку дисперсии σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{o,i} - H_i^T\hat{\bar{\theta}})^2 . \qquad (10.8)$$

Оценка ординаты $y = H^T\bar{\theta}$ гипотетической функции регрессии в любой точке X, очевидно, будет определяться формулой

$$\hat{y} = H^T\hat{\bar{\theta}}, \qquad (10.9)$$

где $H^T = H^T(X)$, а оценка ординаты $y_i = H_i^T\bar{\theta}$ гипотетической функции регрессии в точке наблюдения X_i – формулой

$$\hat{y}_i = H_i^T\hat{\bar{\theta}},$$

где $H_i^T = H^T(X_i)$. Выражение (10.9) представляет собой построенную *эмпирическую* функцию регрессии.

Чаще полученные оценки записывают в другой форме. Рассматривают весь набор данных эксперимента

$$y_{o,1} = H_1^T\bar{\theta} + z_1 = h_1(X_1)\theta_1 + h_2(X_1)\theta_2 + ... + h_m(X_1)\theta_m + z_1,$$

$$y_{o,2} = H_2^T\bar{\theta} + z_2 = h_1(X_2)\theta_1 + h_2(X_2)\theta_2 + ... + h_m(X_2)\theta_m + z_2,$$

……..

$$y_{o,n} = H_n^T\bar{\theta} + z_n = h_1(X_n)\theta_1 + h_2(X_n)\theta_2 + ... + h_m(X_n)\theta_m + z_n .$$

Этот набор данных можно записать в виде

$$Y_o = F\bar{\theta} + Z , \qquad (10.10)$$

где $Z^T = (z_i),\ i = \overline{1,n}$, – вектор ошибок измерений, который распределен по нормальному закону,

$$Z \in N(0, \sigma^2 I),$$

I – единичная матрица, $Y_o^T = (y_{oi}),\ i = \overline{1,n}$, – вектор наблюдений выходной величины объекта, $\bar{\theta}^T = (\theta_j),\ j = \overline{1,m}$, – вектор параметров объекта, $F = (f_{ij}) = (h_j(X_i)),\ i = \overline{1,n},\ j = \overline{1,m}$, – $(n \times m)$-матрица, которая называется

матрицей плана эксперимента. В развернутом виде указанные векторы и матрицы выглядят следующим образом:

$$Y_o = \begin{pmatrix} y_{o1} \\ y_{o2} \\ \vdots \\ y_{on} \end{pmatrix},\ Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix},\ F = \begin{pmatrix} H_1^T \\ H_2^T \\ \vdots \\ H_n^T \end{pmatrix} = \begin{pmatrix} h_1(X_1) & h_2(X_1) & \cdots & h_m(X_1) \\ h_1(X_2) & h_2(X_2) & \cdots & h_m(X_2) \\ \vdots & \vdots & \vdots & \vdots \\ h_1(X_n) & h_2(X_n) & \cdots & h_m(X_n) \end{pmatrix}.$$

Математическое ожидание и ковариационная (дисперсионная) матрица вектора ошибок измерений Z определяются выражениями:

$$E(Z) = 0,\quad cov(Z) = E(ZZ^T) = \sigma^2 I.$$

Легко увидеть, что

$$\sum_{i=1}^{n} H_i H_i^T = F^T F,\qquad \sum_{i=1}^{n} H_i y_{oi} = F^T Y_o,$$

и полученные оценки (10.7), (10.8) запишутся следующим образом:

$$\hat{\bar{\theta}} = (F^T F)^{-1}(F^T Y_o), \tag{10.11}$$

$$\hat{\sigma}^2 = \frac{1}{n}(Y_o - F\hat{\bar{\theta}})^T (Y_o - F\hat{\bar{\theta}}). \tag{10.12}$$

Выражение (10.9) для оценки ординаты гипотетической функции регрессии в любой точке X не изменится:

$$\hat{y} = H^T \hat{\bar{\theta}}.$$

10.2.3 Рекуррентная форма оценок параметров функции регрессии

На практике часто наблюдения $H_i = H_i(X_i)$, $y_{o,i}$ для модели (10.3) поступают последовательно, одно за другим, в реальном времени. В этом случае представляется целесообразным обрабатывать наблюдения по мере их поступления. Получим рекуррентную форму оценок параметров функции регрессии,

позволяющую обрабатывать наблюдения по мере их поступления. Для этого в выражении для оценки (10.7) обозначим

$$A_n = \sum_{i=1}^{n} H_i H_i^T \ , \tag{10.13}$$

$$B_n = \sum_{i=1}^{n} H_i y_{o,i} \ . \tag{10.14}$$

Тогда

$$\hat{\bar{\theta}}_n = A_n^{-1} B_n . \tag{10.15}$$

Поскольку

$$A_n = A_{n-1} + H_n H_n^T , \tag{10.16}$$

$$B_n = B_{n-1} + H_n y_{o,n} , \tag{10.17}$$

то алгоритм (10.16), (10.17), (10.15) представляет собой рекуррентную форму оценок. Статистики A_n, B_n (10.13), (10.14) называются достаточными, поскольку содержат в себе все предыдущие измерения, и достаточны для получения оценки на основе одного нового измерения.

Если обратить матрицу A_n (10.16) как сумму, то получим формулу [17]

$$A_n^{-1} = A_{n-1}^{-1} - A_{n-1}^{-1} H_n (I + H_n^T A_{n-1}^{-1} H_n)^{-1} H_n^T A_{n-1}^{-1} . \tag{10.18}$$

Подставляя (10.18), (10.17) в (10.15), получим рекуррентную формулу в виде

$$\hat{\bar{\theta}}_n = \hat{\bar{\theta}}_{n-1} + A_n^{-1} H_n (y_{o,n} - H_n^T \hat{\bar{\theta}}_{n-1}) .$$

Следует отметить, что полученная рекуррентная форма для оценки может оказаться чувствительной к выбору первоначального приближения $(A_0, B_0, \hat{\bar{\theta}}_0)$, необходимого для организации начала расчетов. В этом случае можно рекомендовать получение сначала оценки по общей формуле (10.7) с использование минимально необходимого числа наблюдений m, а затем переход к рекуррентной формуле.

10.2.4 Свойства оценок параметров функции регрессии

Исследуем свойства полученных оценок. Начнем с оценки $\hat{\bar{\theta}}$ (10.11), которую с учетом выражения (10.10) представим в виде:

$$\hat{\bar{\theta}} = (F^T F)^{-1}(F^T Y_o) = (F^T F)^{-1} F^T (F\bar{\theta} + Z) = \bar{\theta} + (F^T F)^{-1} F^T Z \ . \qquad (10.19)$$

Отсюда получаем, что

$$\hat{\bar{\theta}} - \bar{\theta} = (F^T F)^{-1} F^T Z \, . \qquad (10.20)$$

Найдем математическое ожидание оценки $\hat{\bar{\theta}}$, используя выражение (10.19):

$$E(\hat{\bar{\theta}}) = E(\bar{\theta}) + E((F^T F)^{-1} F^T Z) = \bar{\theta} \, .$$

Мы видим, что оценка $\hat{\bar{\theta}}$ – несмещенная. Найдем также дисперсионную матрицу $D(\hat{\bar{\theta}})$ этой оценки. Поскольку $E(\hat{\bar{\theta}}) = \bar{\theta}$, то вместо (10.20) можно записать:

$$\hat{\bar{\theta}} - E(\hat{\bar{\theta}}) = (F^T F)^{-1} F^T Z \, .$$

Учитывая это выражение, получим

$$D(\hat{\bar{\theta}}) = E((\hat{\bar{\theta}} - E(\hat{\bar{\theta}}))(\hat{\bar{\theta}} - E(\hat{\bar{\theta}}))^T) = E((F^T F)^{-1} F^T Z Z^T F (F^T F)^{-1}) =$$

$$= (F^T F)^{-1} F^T E(ZZ^T) F (F^T F)^{-1} \, .$$

Поскольку $E(Z Z^T) = \sigma^2 I$, то из последнего выражения получаем, что

$$D(\hat{\bar{\theta}}) = \sigma^2 (F^T F)^{-1} \, .$$

Дисперсионная матрица $D(\hat{\bar{\theta}})$ характеризует точность оценки $\hat{\bar{\theta}}$. Для упрощения введем обозначения

$$A = F^T F = (a_{i,j}), \ A^{-1} = (F^T F)^{-1} = (a^{i,j}), \ i, j = \overline{1,m} \, .$$

Тогда $D(\hat{\bar{\theta}}) = \sigma^2 A^{-1}$.

Исследуем оценку $\hat{\bar{\theta}}$ на эффективность, для чего найдем информационную матрицу Фишера (см. раздел 1.5)

$$I(\bar{\theta}) = (I_{i,j}) = -E\left(\frac{\partial^2 \ln L(\bar{\theta})}{\partial\theta_i \partial\theta_j}\right) = -E\left(\frac{d^2 \ln L(\bar{\theta})}{d\bar{\theta}^2}\right).$$

Так как, в силу (10.10),

$$L(\bar{\theta}) = \frac{1}{\sqrt{(2\pi)^n (\sigma^2)^n}} exp\left(-\frac{1}{2\sigma^2}(Y - F\bar{\theta})^T (Y - F\bar{\theta})\right),$$

то

$$\ln L(\bar{\theta}) \sim -\frac{1}{2\sigma^2}(Y - F\bar{\theta})^T (Y - F\bar{\theta}) = \frac{-1}{2\sigma^2}\left(Y^T Y - 2Y^T F\bar{\theta} + \bar{\theta}^T F^T F\bar{\theta}\right).$$

Дифференцируя это выражение по $\bar{\theta}$ два раза, получим

$$\frac{d}{d\bar{\theta}} \ln L(\bar{\theta}) = \frac{1}{\sigma^2} Y^T F - \frac{1}{\sigma^2} F^T F\bar{\theta},$$

$$\frac{d^2}{d\bar{\theta}^2} \ln L(\bar{\theta}) = -\frac{1}{\sigma^2} F^T F.$$

В итоге получаем

$$I(\bar{\theta}) = \frac{1}{\sigma^2} F^T F.$$

Мы видим, что $I^{-1}(\bar{\theta}) = \sigma^2 (F^T F)^{-1} = D(\hat{\bar{\theta}})$, неравенство Рао-Крамера (1.1) превращается в равенство, так что оценка $\hat{\bar{\theta}}$ эффективная.

Найдем теперь математическое ожидание оценки $\hat{\sigma}^2$ (10.12) дисперсии σ^2. Понятно, что

$$\hat{\sigma}^2 = \frac{1}{n} tr\left((Y_o - F\hat{\bar{\theta}})(Y_o - F\hat{\bar{\theta}})^T\right) = \frac{1}{n} tr(C),$$

$$E\left(\hat{\sigma}^2\right) = \frac{1}{n} E(tr(C)),$$

где $C = (Y_o - F\hat{\bar{\theta}})(Y_o - F\hat{\bar{\theta}})^T$, и $tr(C)$ означает след матрицы C. Вычитая и прибавляя в скобках последнего выражения матрицу $F\bar{\theta}$ и учитывая выражения (10.10), (10.20), получим

$$C = \left(Z - F^T (FF^T)^{-1} FZ\right)\left(Z - F^T (FF^T)^{-1} FZ\right)^T =$$

$$= ZZ^T - F^T (FF^T)^{-1} FZZ^T F^T (FF^T)^{-1} F,$$

$$E(C) = \sigma^2 \left(I_n - F^T (FF^T)^{-1} F\right).$$

Поскольку $E(tr(C)) = tr(E(C))$, то

$$E\left(\hat{\sigma}^2\right) = \frac{\sigma^2}{n} tr\left(I_n - F^T (FF^T)^{-1} F\right) = \frac{\sigma^2}{n}\left(tr(I_n) - tr(F^T (FF^T)^{-1} F)\right).$$

Известно, что для двух матриц B, D, таких, что BD и DB существуют, выполняется равенство

$$tr(BD) = tr(DB).$$

Применяя это свойство к матрицам $B = F^T$, $D = (FF^T)^{-1} F$, будем иметь

$$tr(F^T (FF^T)^{-1} F) = tr((FF^T)^{-1} FF^T) = tr(I_m) = m.$$

В итоге получим

$$E\left(\hat{\sigma}^2\right) = \frac{\sigma^2}{n}\left(tr(I_n) - tr(I_m)\right) = \frac{\sigma^2 (n-m)}{n} = \sigma^2 - \frac{\sigma^2 m}{n}.$$

Мы видим, что оценка $\hat{\sigma}^2$ (10.12) смещенная. Однако она асимптотически несмещенная, поскольку ее смещение

$$b(\sigma^2) = -\frac{\sigma^2 m}{n} \xrightarrow[n\to\infty]{} 0.$$

В силу линейности смещения относительно параметра σ^2 его можно устранить. Исправленная оценка

$$\hat{\sigma}_1^2 = \frac{n}{n-m}\hat{\sigma}^2 = \frac{1}{n-m}(Y_o - F\hat{\bar{\theta}})^T (Y_o - F\hat{\bar{\theta}}),$$

где m – количество оцениваемых параметров, является несмещенной.

Наконец, рассмотрим оценку $\hat{y} = H^T \hat{\bar{\theta}}$ (10.9) ординаты $y = H^T \bar{\theta}$ гипотетической функции регрессии. Математическое ожидание этой оценки

$$E(\hat{y}) = E(H^T \hat{\bar{\theta}}) = H^T E(\hat{\bar{\theta}}) = H^T \bar{\theta} = y,$$

то есть совпадает с гипотетическим значением функции регрессии. Следовательно, оценка $\hat{y} = H^T \hat{\bar{\theta}}$ несмещенная. С учетом этого будем иметь

$$\hat{y} - E(\hat{y}) = H^T (\hat{\bar{\theta}} - \bar{\theta}).$$

Найдем дисперсию оценки $\hat{y} = H^T \hat{\bar{\theta}}$.

$$\sigma_{\hat{y}}^2 = D(\hat{y}) = E((\hat{y} - E(\hat{y}))(\hat{y} - E(\hat{y}))^T) = E(H^T (\hat{\bar{\theta}} - \bar{\theta})(\hat{\bar{\theta}} - \bar{\theta})^T H) =$$

$$= H^T E((\hat{\bar{\theta}} - \bar{\theta})(\hat{\bar{\theta}} - \bar{\theta})^T) H = \sigma^2 H^T (F^T F)^{-1} H.$$

Эта дисперсия характеризует точность оценки $\hat{y} = H^T \hat{\bar{\theta}}$.

В заключение отметим, что оценки $\hat{\bar{\theta}}$ и $\hat{\sigma}^2$ независимы. Следовательно, независимы также оценки $\hat{\bar{\theta}}$ и $\hat{\sigma}_1^2$, $\hat{y}$ и $\hat{\sigma}_1^2$.

10.2.5 Распределения оценок параметров функции регрессии

Рассмотрим распределения оценок. Поскольку оценка $\hat{\bar{\theta}}$ (10.11) является линейной функцией вектора Y, имеющего нормальное распределение, то она также будет иметь нормальное распределение. Параметры этого распределения (математическое ожидание и Дисперсионная матрица) были получены нами в предыдущем разделе. Следовательно, мы можем записать, что

$$\hat{\bar{\theta}} \in N_m(\bar{\theta}, \sigma^2 (F^T F)^{-1}). \qquad (10.21)$$

Поскольку оценка $\hat{y}$ является линейной функцией вектора $\hat{\bar{\theta}}$ с нормальным распределением, то она также распределена по нормальному закону. Параметры этого закона (математическое ожидание и дисперсию) мы получили в предыдущем разделе. Следовательно

$$\hat{y} = H^T \hat{\bar{\theta}} \in N(H^T \theta, \sigma^2 H^T A^{-1} H).$$

Без доказательства укажем, что статистика v имеет распределение хи-квадрат с $n - m$ степенями свободы:

$$v = \frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-m)\hat{\sigma}_1^2}{\sigma^2} \in H_{n-m}.$$

Учитывая эти распределения и независимость оценок $\hat{\bar{\theta}}$ и $\hat{y}$ от статистики v, мы можем стандартным образом (см. раздел 3.4) получить другие статистики и их законы распределения. Учитывая распределение (10.21) вектора $\hat{\bar{\theta}}$, можно записать, что

$$u_i = \frac{\hat{\theta}_i - \theta_i}{\sqrt{\sigma^2 a^{ii}}} \in N(0,1), \quad i = \overline{1,m}.$$

Тогда

$$t_i = \frac{u_i}{\sqrt{v}}\sqrt{n-m} = \frac{\hat{\theta}_i - \theta_i}{\sqrt{n\hat{\sigma}^2 a^{ii}}}\sqrt{n-m} = \frac{\hat{\theta}_i - \theta_i}{\sqrt{\hat{\sigma}_1^2 a^{ii}}} \in T_1(n-m), \quad i = \overline{1,m}.$$

Статистики t_i, v используются для построения доверительных интервалов для θ_i, σ^2 соответственно и проверки гипотез об этих параметрах.

Используя распределение оценки $\hat{y}$, можно записать, что

$$u_y = \frac{\left(\hat{y} - H^T\bar{\theta}\right)}{\sqrt{\sigma^2 H^T A^{-1} H}} \in N(0,1).$$

Тогда

$$t_y = \frac{u_y}{\sqrt{v}}\sqrt{n-m} = \frac{\left(\hat{y} - y\right)\sqrt{n-m}}{\sqrt{H^T A^{-1} H}\sqrt{n\hat{\sigma}^2}} = \frac{(\hat{y} - y)}{\sqrt{H^T A^{-1} H \hat{\sigma}_1^2}} \in T_1(n-m).$$

Статистика t_y может применяться для построения доверительного интервала для гипотетического значения ординаты $y = H^T\bar{\theta}$ функции регрессии и проверки гипотез относительно $y = H^T\bar{\theta}$.

10.2.6 Построение доверительных интервалов и проверка гипотез об отдельных параметрах функции регрессии

Статистики, полученные в предыдущем разделе, позволяют строить доверительные интервалы для неизвестных параметров и характеристик по методике, изложенной в разделе 4.2. Приведем эти интервалы. На основе статистики t_i получаем доверительный интервал для параметра θ_i, $i=\overline{1,m}$:

$$\widehat{\theta}_i - t_{\frac{1-\gamma}{2}}\sqrt{\widehat{\sigma}_1^2 a^{i,i}} < \theta_i < \widehat{\theta}_i + t_{\frac{1-\gamma}{2}}\sqrt{\widehat{\sigma}_1^2 a^{i,i}}\ .$$

На основе статистики t_y получаем доверительный интервал для гипотетического значения ординаты $y = H^T\overline{\theta}$ функции регрессии:

$$\hat{y} - t_{\frac{1-\gamma}{2}}\sqrt{\widehat{\sigma}_1^2 H^T A^{-1} H} < y < \hat{y} + t_{\frac{1-\gamma}{2}}\sqrt{\widehat{\sigma}_1^2 H^T A^{-1} H}\ .$$

В этих выражениях $t_{\frac{1-\gamma}{2}}$ – $-100\dfrac{1-\gamma}{2}$-процентное отклонение распределения $T_1(n-m)$ (Стьюдента с $n-m$ степенями свободы).

На основе статистики v получаем доверительный интервал для гипотетической дисперсии σ^2:

$$\frac{(n-m)\widehat{\sigma}_1^2}{v_{\frac{1-\gamma}{2}}} < \sigma^2 < \frac{(n-m)\widehat{\sigma}_1^2}{v_{\frac{1+\gamma}{2}}},$$

где $v_{\frac{1-\gamma}{2}}$, $v_{\frac{1+\gamma}{2}}$ – $100\dfrac{1-\gamma}{2}$ и $100\dfrac{1+\gamma}{2}$-процентные отклонения распределения $H_1(n-m)$ (хи-квадрат с $n-m$ степенями свободы).

Приведенные в разделе 10.2.4 распределения позволяют выполнять проверку гипотез о параметрах θ_i, $i=\overline{1,m}$, функции регрессии. Например, можно проверять гипотезу $\{H_0 : \theta_i = \theta_{i,0};\ \ H_1 : \theta_i \neq \theta_{i,0}\}$, где $\theta_{i,0}$ – некоторое фиксирован-

ное число. В случае $\theta_{i,0} = 0$ эта гипотеза называется гипотезой о значимости коэффициента θ_i, $i = \overline{1,m}$. Приведенная гипотеза проверяется на основе статистики t_i. Критерий для проверки этой гипотезы имеет вид:

$$P(|t_i| > t_{\frac{\alpha}{2}}) = \alpha,$$

где α – уровень значимости. Гипотеза проверяется следующим образом. Выбрав α, например, $\alpha = 0{,}05$, определяем предел значимости $t_{\frac{\alpha}{2}}$, используя для этого таблицу процентных отклонений распределения Стьюдента, выбрав в ней $n - m$ степеней свободы и $\alpha = 0{,}05$. Затем по формуле для статистики t_i при $\theta_i = \theta_{i,0}$ определяем эмпирическое значение статистики $t_{i,э}$. Если окажется, что $|t_{i,э}| > t_{\frac{\alpha}{2}}$, то проверяемая гипотеза $H_0 : \theta_i = \theta_{i,0}$ отклоняется. Если гипотеза *о значимости* коэффициента θ_i, $i = \overline{1,m}$, (при $\theta_{i,0} = 0$) была отклонена (коэффициент признан незначимым), то в функции регрессии им можно пренебречь, то есть можно считать его равным нулю.

10.2.7 Проверка гипотезы о линейности функции регрессии

Пусть требуется проверить гипотезу о том, что все параметры линейной функции регрессии (10.2), кроме свободного члена, равны нулю, то есть гипотезу вида $\{H_0, H_1\}$, где

$$H_0 : \theta_2 = \theta_3 = ... = \theta_m = 0, \qquad (10.22)$$

а альтернатива H_1 состоит в том, что хотя бы одно из указанных равенств не выполняется. Такая гипотеза называется гипотезой о существовании линейной стохастической связи между выходной переменной y и входными переменны-

ми $h_1(X),...,h_m(X)$. Для проверки такой гипотезы применяются методы *дисперсионного анализа,* основанные на анализе так называемых сумм квадратов.

Величина R_0^2, равная сумме квадратов отклонений наблюденных значений выходной величины y_{oi} от значений $\hat{y}_i = H_i^T \hat{\theta}$, предсказанных (полученных) по модели, называется остаточной суммой квадратов:

$$R_0^2 = \sum_{i=1}^{n} (y_{oi} - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_{oi} - H_i^T \hat{\theta})^2 .$$

Величина R_1^2, равная сумме квадратов отклонений наблюденных значений выходной величины y_{oi} от среднего арифметического этих значений

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_{oi}$$

называется общей или полной суммой квадратов:

$$R_1^2 = \sum_{i=1}^{n} (y_{oi} - \bar{y})^2 .$$

Величина R^2, равная сумме квадратов отклонений значений $\hat{y}_i = H_i^T \hat{\theta}$, предсказанных (полученных) по модели, от среднего арифметического наблюденных значений $\bar{y}$, называется суммой квадратов, обусловленной регрессией:

$$R^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

Относительно указанных сумм квадратов справедлива следующая теорема.

Теорема 10.1. Если верна гипотеза H_0 (10.22), то

1) выполняется равенство

$$R_1^2 = R_0^2 + R^2,$$

или, в развернутом виде,

$$\sum_{i=1}^{n} (y_{oi} - \bar{y})^2 = \sum_{i=1}^{n} (y_{oi} - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 ;$$

2) величины R_0^2 / σ^2 и $R^2 / \sigma^2 = (R_1^2 - R_0^2) / \sigma^2$ независимы и имеют распределения хи-квадрат с $n-m$ и $m-1$ степенями соответственно;

2) отношение

$$F = \frac{R_1^2 - R_0^2}{m-1} \Big/ \frac{R_0^2}{n-m} \tag{10.23}$$

имеет F-распределение (Фишера) с $m-1,\ n-m$ степенями свободы,

$$F = \frac{R_1^2 - R_0^2}{m-1} \Big/ \frac{R_0^2}{n-m} \in F_1(m-1, n-m).$$

Для доказательства выполним следующие очевидные преобразования остаточной суммы квадратов:

$$R_0^2 = \sum_{i=1}^{n} (y_{oi} - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_{oi} - H_i^T \hat{\bar{\theta}})^2 = \sum_{i=1}^{n} (y_{oi}^2 - 2 y_{oi} H_i^T \hat{\bar{\theta}} + \hat{\bar{\theta}}^T H_i H_i^T \hat{\bar{\theta}}) =$$

$$= \sum_{i=1}^{n} y_{oi}{}^2 - 2 \sum_{i=1}^{n} y_{oi} H_i^T \hat{\bar{\theta}} + (\sum_{i=1}^{n} \hat{\bar{\theta}}^T H_i H_i^T) \hat{\bar{\theta}}.$$

Из первого нормального уравнения (10.5) мы имеем

$$\sum_{i=1}^{n} \hat{\bar{\theta}}^T H_i H_i^T = \sum_{i=1}^{n} y_{oi} H_i^T .$$

Подставляя это выражение в предыдущее, получим

$$R_0^2 = \sum_{i=1}^{n} y_{oi}{}^2 - 2 \sum_{i=1}^{n} y_{oi} H_i^T \hat{\bar{\theta}} + \sum_{i=1}^{n} y_{oi} H_i^T \hat{\bar{\theta}} = \sum_{i=1}^{n} y_{oi}^2 - \sum_{i=1}^{n} y_{oi} H_i^T \hat{\bar{\theta}}.$$

Прибавим и вычтем в правой части этого выражения величину $n\bar{y}^2$:

$$R_0^2 = (\sum_{i=1}^{n} y_{oi}^2 - n\bar{y}^2) - (\sum_{i=1}^{n} y_{oi} H_i^T \hat{\bar{\theta}} - n\bar{y}^2). \tag{10.24}$$

Докажем теперь, что первое слагаемое правой части этого выражения равно R_1^2, а второе слагаемое равно R^2, то есть докажем, что

$$R_0^2 = R_1^2 - R^2. \tag{10.25}$$

Для этого преобразуем выражение для R_1^2:

$$R_1^2 = \sum_{i=1}^{n} (y_{oi} - \bar{y})^2 = \sum_{i=1}^{n} y_{oi}^2 - 2\sum_{i=1}^{n} y_{oi}\bar{y} + \sum_{i-1}^{n} \bar{y}^2 =$$

$$= \sum_{i=1}^{n} y_{oi}^2 - 2n\bar{y}^2 + n\bar{y}^2 = \sum_{i=1}^{n} y_{oi}^2 - n\bar{y}^2 .$$

В результате мы получили первое слагаемое в правой части равенства (10.24). Рассмотрим теперь выражение для R^2:

$$R^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \hat{y}_i^2 - 2\sum_{i=1}^{n} \hat{y}_i\bar{y} + \sum_{i=1}^{n} \bar{y}^2 .$$

Если выполняются равенства (10.22), то предсказанное по уравнению регрессии значение $\hat{y}_i$ оказывается равным

$$\hat{y}_i = \sum_{i=1}^{n} \hat{\theta}_i h_i(X) = \hat{\theta}_1 ,$$

откуда

$$\sum_{i=1}^{n} \hat{y}_i = n\hat{\theta}_1 .$$

С другой стороны, из первого нормального уравнения при условии выполнения равенств (10.22) получим:

$$\sum_{i=1}^{n} y_{oi} = n\hat{\theta}_1 .$$

В итоге мы получили, что $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_{oi}$, или

$$\bar{\hat{y}} = \bar{y},$$

где $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$. Вернемся теперь к преобразованию выражения (10.25) для R^2 и учтем предыдущее равенство. В результате получим

$$R^2 = \sum_{i=1}^{n} \hat{y}_i^2 - 2\sum_{i=1}^{n} \hat{y}_i\bar{y} + \sum_{i=1}^{n} \bar{y}^2 = \sum_{i=1}^{n} y_{oi}^2 - 2n\bar{\hat{y}}\bar{y} + n\bar{y}^2 =$$

$$= \sum_{i=1}^{n} y_{oi}^2 - 2n\bar{y}^2 + n\bar{y}^2 = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2 . \qquad (10.26)$$

Далее, поскольку

$$\sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} y_{oi}^T y_{oi} = \hat{\bar{\theta}}^T (\sum_{i=1}^{n} H_i H_i^T) \hat{\bar{\theta}}$$

и

$$\hat{\bar{\theta}} = (\sum_{i=1}^{n} H_i H_i^T)^{-1} (\sum_{i=1}^{n} H_i y_{oi}),$$

то

$$\sum_{i=1}^{n} \hat{y}_i^2 = \hat{\bar{\theta}}^T (\sum_{i=1}^{n} H_i H_i^T) \hat{\bar{\theta}} =$$

$$= (\sum_{i=1}^{n} H_i y_{oi})^T (\sum_{i=1}^{n} H_i H_i^T)^{-1} (\sum_{i=1}^{n} H_i H_i^T) (\sum_{i=1}^{n} H_i H_i^T)^{-1} (\sum_{i=1}^{n} H_i y_{oi}) =$$

$$= (\sum_{i=1}^{n} H_i y_{oi})^T (\sum_{i=1}^{n} H_i H_i^T)^{-1} (\sum_{i=1}^{n} H_i y_{oi}) = \sum_{i=1}^{n} y_{oi} H_i^T \hat{\bar{\theta}} .$$

Подставляя полученное выражение для $\sum_{i=1}^{n} \hat{y}_i^2$ в выражение (10.26) для R^2, мы получим, что R^2 совпадает со вторым слагаемым в (10.24):

$$R^2 = \sum_{i=1}^{n} y_{oi} H_i^T \hat{\bar{\theta}} - n\bar{y}^2 .$$

Итак, равенство (10.24) доказано.

Утверждение 2) теоремы о независимости величин R_0^2 / σ^2 и $R^2 / \sigma^2 = (R_1^2 - R_0^2) / \sigma^2$ и их законах распределения примем без доказательства. Ограничимся лишь пояснениями о подсчете чисел степеней свободы этих величин. Остаточная сумма квадратов R_0^2 имеет $n - m$ степеней свободы, поскольку она определяется суммой квадратов n случайных величин $y_{o1}, y_{o2}, ..., y_{on}$, на которые наложено m линейных связей посредством зависимостей $\hat{y}_i$, определяемых через оценки m параметров. С полной суммой квад-

ратов R_1^2 связано $n-1$ степеней свободы, поскольку в нее входит n наблюдений $y_{o1}, y_{o2}, ..., y_{on}$, на которые наложена одна линейная связь $\bar{y}$. С суммой квадратов R^2, обусловленной регрессией, связано $m-1$ степеней свободы, поскольку $\hat{y}_i$ определяет m линейных связей между наблюдениями посредством коэффициентов $\theta_1, \theta_2, ..., \theta_m$, определенных по тем же наблюдениям. Кроме того, $\bar{y}$ определяет одну линейную связь между ними.

Утверждение 3) теоремы очевидно, поскольку оно содержит определение случайной величины, имеющей F-распределение Фишера.

Критерий для проверки гипотезы $\{H_0, H_1\}$, где H_0 определяется выражением (10.22), имеет вид:

$$P(F > F_\alpha) = \alpha .$$

На основании приведенной теоремы сформулированная гипотеза проверяется следующим образом. Задаемся уровнем значимости α и по таблице процентных отклонений F-распределения с $m-1, n-m$ степенями свободы определяем уровень значимости F_α. Затем по формуле (10.23) вычисляем эмпирическое значение $F_э$ статистики F. Если окажется, что $F_э > F_\alpha$, то проверяемая гипотеза H_0 отклоняется. Это будет означать, что линейная стохастическая зависимость между выходной переменной y и входными переменными $h_1(X), ..., h_m(X)$ существует. Если же $F_э \le F_\alpha$, то гипотеза H_0 принимается. Это будет означать, что линейная стохастическая зависимость между выходной переменной y и входными переменными $h_1(X), ..., h_m(X)$ слаба или отсутствует вовсе. Такой результат возможен по следующим причинам. Во-первых, в модель не включены некоторые из сильно влияющих факторов $h_i(X)$, не замеченных экспериментатором. Во-вторых, в модель включены все существенно влияющие факторы, но структура модели выбрана неправильно, например, модель является существенно нелинейной относительно факторов. Принятие гипотезы H_0 является достаточным основанием для того, чтобы отказаться от

выбранной модели с тем, чтобы попытаться построить новую модель. Если все попытки завершаются принятием гипотезы H_0, то остается сделать вывод, что зависимость между y и $h_1(X),...,h_m(X)$ действительно отсутствует, как линейная, так и нелинейная.

10.2.8 Проверка адекватности регрессионной модели

Важным вопросом после получения оценок параметров функции регрессии является проверка пригодности полученной модели. Такая проверка получила название *проверки адекватности модели*. Проверка адекватности модели – это проверка того, что модель постулирована априори верно, что выбранная модель соответствует действительности. Эта задача формулируется как задача проверки вполне определенной двухальтернативной гипотезы $\{H_0, H_1\}$ [8]. Чтобы сформулировать эту гипотезу, обратимся к следующим рассуждениям. Как указывалось выше, величина

$$\hat{\sigma}_1^2 = \frac{1}{n-m}\sum_{i=1}^{n}(y_{oi} - \hat{y}_i)^2 \qquad (10.27)$$

является несмещенной оценкой дисперсии σ^2 гипотетической модели, а величина

$$v = \frac{(n-m)\hat{\sigma}_1^2}{\sigma^2}$$

имеет распределение хи-квадрат с $n-m$ степенями свободы.

Выполнив k дополнительных опыта в некоторой точке X с результатами $y_{\partial 1}, y_{\partial 2},..., y_{\partial k}$, можно получить другую независимую оценку $\hat{\sigma}_\partial^2$ дисперсии модели:

$$\hat{\sigma}_\partial^2 = \frac{1}{k-1}\sum_{i=1}^{k}(y_{\partial i} - \bar{y}_\partial)^2, \qquad (10.28)$$

где

$$\bar{y}_{\partial} = \frac{1}{k}\sum_{i=1}^{k} y_{\partial i} .$$

Наблюдения $y_{\partial 1}, y_{\partial 2}, ..., y_{\partial k}$ не должны использоваться при получении оценок $\hat{\bar{\theta}}$, $\hat{\sigma}_1^2$. Оценка $\hat{\sigma}_{\partial}^2$ является несмещенной оценкой дисперсии $\sigma_{м}^2$ любой, а значит и действительной, модели с однородной дисперсией, а величина

$$w = \frac{(k-1)\hat{\sigma}_{\partial}^2}{\sigma_{м}^2}$$

имеет распределение хи-квадрат с $k-1$ степенями свободы. Проверка адекватности модели состоит в проверке гипотезы о равенстве гипотетической и действительной дисперсий:

$$\{ H_0 : \sigma^2 = \sigma_{м}^2 ;\ H_1 : \sigma^2 > \sigma_{м}^2 \}.$$

Критерий для проверки этой гипотезы основывается на сравнении оценок этих дисперсий (10.27), (10.28). Рассматривается отношение

$$F = \frac{v}{n-m} \Big/ \frac{w}{k-1} = \frac{\hat{\sigma}_1^2 \sigma_{м}^2}{\sigma^2 \sigma_{\partial}^2},$$

которое при истинности проверяемой гипотезы $H_0 : \sigma_{м}^2 = \sigma^2$ принимает вид

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_{\partial}^2} \qquad (10.29)$$

и имеет F-распределение Фишера с $n-m, k-1$ степенями свободы. Критерий проверки гипотезы имеет вид

$$P(F > F_{\alpha}) = \alpha .$$

Гипотеза проверяется следующим образом. После оценивания регрессионных коэффициентов вычисляют $\hat{\sigma}_1^2$ по формуле (10.27). Проводят k дополнительных опытов и по формуле (10.28) получают независимую оценку $\hat{\sigma}_{\partial}^2$ дисперсии $\sigma_{м}^2$ случайной ошибки. Вычисляют эмпирическое значение $F_{э}$ дисперсионного

отношения F по формуле (10.29). Выбирают уровень значимости α, например, $\alpha = 0{,}05$, и по таблицам процентных отклонений F-распределения Фишера с $n-m, k-1$ степенями свободы находят предел значимости F_α. Сравнивают величины $F_э$ и F_α и делают один из следующих выводов. Если $F_э \le F_\alpha$, то гипотеза $H_0 : \sigma^2 = \sigma_м^2$ принимается и модель считается адекватной, поскольку оценка $\hat{\sigma}_1^2$ оказалась соизмеримой с оценкой $\hat{\sigma}_д^2$. Если $F_э > F_\alpha$, то гипотеза $H_0 : \sigma^2 = \sigma_м^2$ отклоняется и модель считается неадекватной, поскольку оценка $\hat{\sigma}_1^2$ оказалась значимо больше оценки $\hat{\sigma}_д^2$. В увеличении $\hat{\sigma}_1^2$ сыграла свою роль неадекватность модели. При констатации неадекватности модели надо изменить ее структуру и заново обработать данные. Например, если полином первой степени оказался неадекватным, надо перейти к полиному второй степени.

10.2.9 Случай простой линейной регрессии

Представляется полезным рассмотреть случай простой линейной регрессии как частный случай множественной регрессии. Это позволит полнее усвоить технику векторно-матричных обозначений и преобразований множественного регрессионного анализа.

В случае простой линейной регрессии $\varphi(x) = a + bx$ модель наблюдения описывается в виде

$$y_i = H_i^T \theta + \varepsilon_i,\ i = \overline{1,n},$$

где

$$H_i^T = (1, x_i),\ \theta^T = (a, b).$$

Напомним, что оценка вектора параметров имеет вид (см. (10.7))

$$\hat{\theta} = (\sum_{i=1}^{n} H_i H_i^T)^{-1} (\sum_{i=1}^{n} H_i y_i)$$

или

$$\hat{\bar{\theta}} = (F^T F)^{-1} F^T Y,$$

где в случае простой линейной регрессии

$$F^T F = A = \sum_{i=1}^{n} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} =$$

$$= n \begin{bmatrix} 1 & \frac{1}{n}\sum x_i \\ \frac{1}{n}\sum x_i & \frac{1}{n}\sum x_i^2 \end{bmatrix} = n \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \bar{s}_x^2 + \bar{x}^2 \end{bmatrix},$$

$$\bar{s}_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2,$$

$$|A| = n^2 (\bar{s}_x^2 + \bar{x}^2 - \bar{x}^2) = n^2 \bar{s}_x^2,$$

$$A^{-1} = \frac{1}{n\bar{s}_x^2} \begin{bmatrix} \bar{s}_x^2 + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} = \begin{bmatrix} \frac{\bar{s}_x^2 + \bar{x}^2}{n\bar{s}_x^2} & -\frac{\bar{x}}{n\bar{s}_x^2} \\ -\frac{\bar{x}}{n\bar{s}_x^2} & \frac{1}{n\bar{s}_x^2} \end{bmatrix} = \begin{bmatrix} a^{1,1} & a^{1,2} \\ a^{2,1} & a^{2,2} \end{bmatrix},$$

$$\sum_{i=1}^{n} H_i y_i = F^T Y = \sum_{i=1}^{n} \begin{bmatrix} 1 \\ x_i \end{bmatrix} y_i = \sum \begin{bmatrix} y_i \\ y_i x_i \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \end{bmatrix} = n \begin{bmatrix} \bar{y} \\ \bar{s}_{xy} + \bar{x}\,\bar{y} \end{bmatrix},$$

$$\bar{s}_{xy} = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}_i).$$

Оценка в векторной форме имеет вид

$$\hat{\bar{\theta}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \frac{\bar{s}_x^2 + \bar{x}^2}{\bar{s}_x^2} \bar{y} - \frac{\bar{x}}{\bar{s}_x^2}(\bar{s}_{xy} + \bar{x}\,\bar{y}) \\ -\frac{\bar{x}^2}{\bar{s}_x^2} \bar{y} + \frac{1}{\bar{s}_x^2}(\bar{s}_{xy} + \bar{x}\,\bar{y}) \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{\bar{s}_{xy}}{\bar{s}_x^2} \bar{x} \\ \frac{\bar{s}_{xy}}{\bar{s}_x^2} \end{bmatrix},$$

откуда получаем оценки отдельных параметров простой линейной регрессии

$$\hat{a} = \frac{\bar{s}_x^2 \bar{y} + \bar{x}^2 \bar{y} - \bar{x}\bar{s}_{xy} - \bar{x}^2 \bar{y}}{\bar{s}_x^2} = \bar{y} - \frac{\bar{s}_{xy}}{\bar{s}_x^2} \bar{x},$$

$$\hat{b} = \frac{-\bar{x}\,\bar{y} + \bar{s}_{xy} + \bar{x}\,\bar{y}}{\bar{s}_x^2} = \frac{\bar{s}_{xy}}{\bar{s}_x^2}.$$

Мы получили те же оценки, что и в разделе 10.1 путем привлечения метода наименьших квадратов и еще раньше в разделе 9.1 путем привлечения выборочного метода. Следовательно, выборочный метод, метод наименьших квадратов и метод максимума правдоподобия в данном случае оказываются эквивалентными.

Продолжим анализ случая простой линейной регрессии исходя из результатов для множественной регрессии, учитывая, что теперь $\hat{\theta}_1 = \hat{a}$, $\hat{\theta}_2 = \hat{b}$. Найдем оценки дисперсии исходя из общих формул (см. (10.8))

$$\hat{\sigma}_1^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - H_i^T \hat{\bar{\theta}})^2,$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - H_i^T \hat{\bar{\theta}})^2.$$

Учитывая выражения для H_i^T и $\hat{\bar{\theta}}$, будем иметь

$$H_i^T \hat{\bar{\theta}} = \bar{y} - \frac{\bar{s}_{xy}}{\bar{s}_x^2} \bar{x} + \frac{\bar{s}_{xy}}{\bar{s}_x^2} x_i = \bar{y} + \frac{\bar{s}_{xy}}{\bar{s}_x^2} (x_i - \bar{x}),$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - H_i^T \hat{\bar{\theta}})^2 =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \frac{\bar{s}_{xy}}{\bar{s}_x^2} (x_i - \bar{x}) \right]^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\dot{y}_i^2 - 2\dot{x}_i \dot{y}_i \frac{\bar{s}_{xy}}{\bar{s}_x^2} + \frac{\bar{s}_{xy}^2}{(\bar{s}_x^2)^2} \dot{x}_i^2 \right] =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \dot{y}_i^2 - \frac{2}{n} \sum_{i=1}^{n} \dot{x}_i \dot{y}_i \frac{\bar{s}_{xy}}{\bar{s}_x^2} + \frac{1}{n} \sum_{i=1}^{n} \dot{x}_i^2 \frac{\bar{s}_{xy}^2}{(\bar{s}_x^2)^2} =$$

$$= \bar{s}_y^2 - 2\bar{s}_{xy} \frac{\bar{s}_{xy}}{\bar{s}_x^2} + \bar{s}_x^2 \frac{\bar{s}_{xy}^2}{(\bar{s}_x^2)^2} = \bar{s}_y^2 - \frac{\bar{s}_{xy}^2}{\bar{s}_x^2} = \bar{s}_y^2 (1 - \bar{r}_{xy}^2) = \hat{\sigma}^2,$$

где

$$\dot{x}_i = x_i - \bar{x},$$

$$\dot{y}_i = y_i - \bar{y},$$

$$\bar{r}_{xy} = \frac{\bar{s}_{xy}}{\bar{s}_x \bar{s}_y}.$$

Мы получили то же выражение для оценки $\hat{\sigma}^2$, что и в разделе 10.1. Эта оценка является смещенной. Несмещенной будет оценка $\hat{\sigma}_1^2$:

$$\hat{\sigma}_1^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{n}{n-2}\bar{s}_y^2(1-\bar{r}_{xy}^2).$$

Далее, учитывая, что

$$a^{1,1} = \frac{\bar{s}_{xy}^2 + \bar{x}^2}{n\bar{s}_x^2},$$

$$a^{2,2} = \frac{1}{n\bar{s}_x^2},$$

Получим выражения для t-статистик коэффициентов a и b из общей формулы

$$t_i = \frac{\hat{\theta}_i - \theta_i}{\sqrt{\hat{\sigma}_1^2 a^{i,i}}}.$$

Будем иметь

$$t_1 = t_a = \frac{\hat{a} - a}{\sqrt{\frac{n}{n-2}\bar{s}_y^2(1-\bar{r}_{xy}^2)\frac{\bar{s}_x^2 + \bar{x}^2}{n\bar{s}_x^2}}} = \frac{(\hat{a} - a)\bar{s}_x\sqrt{n-2}}{\sqrt{\bar{s}_y^2(1-\bar{r}_{xy}^2)(\bar{s}_x^2 + \bar{x}^2)}} =$$

$$= \frac{(\hat{a} - a)\bar{s}_x\sqrt{n-2}}{\hat{\sigma}\sqrt{\bar{x}^2 + \bar{s}_x^2}} \in T_{n-2},$$

$$t_2 = t_b = \frac{\hat{b} - b}{\sqrt{\frac{n}{n-2}\bar{s}_y^2(1-\bar{r}_{xy}^2)\frac{1}{n\bar{s}_x^2}}} = \frac{(\hat{b} - b)\bar{s}_x}{\hat{\sigma}_1}\sqrt{n} =$$

$$= \frac{(\hat{b} - b)\bar{s}_x}{\hat{\sigma}_1}\sqrt{n} = \frac{(\hat{b} - b)\bar{s}_x}{\hat{\sigma}}\sqrt{n-2} \in T_{n-2}.$$

Выражение для статистики t_b такое же, как и в разделе 10.1.

ОГЛАВЛЕНИЕ