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

«Допустить к защите»_____

Заведующий кафедрой
д.ф.-м.н., профессор С.М. Ермаков

Учебное пособие

Алексеева Нина Петровна

Прикладная статистика.
Часть 2. Многомерные методы

Санкт-Петербург – 2020

Оглавление

1. Множественная регрессия

Пусть имеются случайные величины $\xi_1, \ldots, \xi_n$ с совместной плотностью $f(x_1, x_2, \ldots, x_n)$. Регрессией или условным математическим ожиданием случайной величины ξ_1 при фиксированных значениях случайных величин $\xi_i,\ i = 2, \ldots, n$, называется функция, зависящая от переменных $x_2, \ldots, x_n$, вида

$$E(\xi_1|\xi_2 = x_2, \ldots, \xi_n = x_n) = m_1(x_2, \ldots, x_n) = \frac{\int\limits_{-\infty}^{\infty} x_1 f(x_1, \ldots, x_n)dx_1}{\int\limits_{-\infty}^{\infty} f(x_1, \ldots, x_n)dx_1}$$

Поверхность регрессии описывается уравнением $x_1 = m_1(x_2, \ldots, x_n)$.

1.1. Линейная среднеквадратичная регрессия

Для упрощения выводов будем считать $E\xi_i = 0$. Формулы для произвольного центра тяжести будут получаться простой заменой ξ_i на $\xi_i - m_i$. Среднеквадратическая регрессия величины ξ_1 относительно $\xi_2, \ldots, \xi_n$ определяется как гиперплоскость

$$\xi_1 = \beta_{12}\xi_2 + \ldots + \beta_{1n}\xi_n, \tag{1}$$

дающая наилучшую аппроксимацию математического ожидания квадрата разности

$$E(\xi_1 - \beta_{12}\xi_2 - \ldots - \beta_{1n}\xi_n)^2 \to min, \tag{2}$$

т.е. правая часть (1) является наилучшей линейной оценкой ξ_1 величинами $\xi_2, \ldots, \xi_n$ в смысле минимума (2). Коэффициенты β_{1i} называются *частными коэффициентами регрессии.*

Для их вычисления продифференцируем (2) по β_{1i}.

$$\begin{cases} \frac{d}{d\beta_{12}}: & -2E\xi_2(\xi_1 - \beta_{12}\xi_2 - \ldots - \beta_{1n}\xi_n) = 0 \\ & \ldots \\ \frac{d}{d\beta_{1n}}: & -2E\xi_n(\xi_1 - \beta_{12}\xi_2 - \ldots - \beta_{1n}\xi_n) = 0 \end{cases}$$

Это равносильно при $\lambda_{ij} = E\xi_i\xi_j$ системе уравнений

$$\begin{cases} \lambda_{22}\beta_{12} + \ldots + \lambda_{2n}\beta_{1n} = \lambda_{21} \\ \ldots \\ \lambda_{n2}\beta_{12} + \ldots + \lambda_{nn}\beta_{1n} = \lambda_{n1} \end{cases} \tag{3}$$

Определитель этой системы равен алгебраическому дополнению Λ_{11} матрицы вторых моментов

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2n} \\ & & \ldots & \\ \lambda_{n1} & \lambda_{n2} & \ldots & \lambda_{nn} \end{bmatrix}$$

По теореме Крамера, β_{12} равно отношению определителя

$$\begin{vmatrix} \lambda_{21} & \lambda_{23} & \ldots & \lambda_{2n} \\ & & \ldots & \\ \lambda_{n1} & \lambda_{n3} & \ldots & \lambda_{nn} \end{vmatrix},$$

который равен $-\Lambda_{12}$, к определителю Λ_{11}, т.е. $\beta_{12} = -\Lambda_{12}/\Lambda_{11}$. Коэффициент β_{13} равен отношению определителя

$$\begin{vmatrix} \lambda_{22} & \lambda_{21} & \lambda_{24} & \ldots & \lambda_{2n} \\ & & & \ldots & \\ \lambda_{n2} & \lambda_{n1} & \lambda_{n4} & \ldots & \lambda_{nn} \end{vmatrix}, \tag{4}$$

полученного в результате замены в матрице Λ_{11} второго столбца на столбец свободных членов системы (3), к определителю Λ_{11}. Поменяем местами

первые два столбца. Определитель (4) равен

$$-\begin{vmatrix} \lambda_{21} & \lambda_{22} & \lambda_{24} & \dots & \lambda_{2n} \\ & & & \dots & \\ \lambda_{n1} & \lambda_{n2} & \lambda_{n4} & \dots & \lambda_{nn} \end{vmatrix} = -\Lambda_{13}, \tag{5}$$

т.е.

$$\beta_{13} = -\frac{\Lambda_{13}}{\Lambda_{11}}.$$

Аналогично получаем, что

$$\beta_{1k} = -\frac{\Lambda_{1k}}{\Lambda_{11}}. \tag{6}$$

Простой перестановкой индексов получаем

$$\beta_{ik} = -\frac{\Lambda_{ik}}{\Lambda_{ii}}. \tag{7}$$

1.2. Остатки и остаточная дисперсия

Ограничимся рассмотрением случая, когда $|\Lambda_{11}| \neq 0$ и все $E\xi_i = 0$. Случайная величина вида

$$\eta_{1\cdot 23\dots n} = \xi_1 - \beta_{12}\xi_2 - \dots - \beta_{1n}\xi_n \tag{8}$$

называется остатком величины ξ_1 относительно $\xi_2, \dots, \xi_n$. Покажем, что *остаток не коррелирован ни с одной из "вычитаемых"случайных величин.* Из (6) и из того, что $1 = \Lambda_{11}/\Lambda_{11}$ имеем выражение остатка через линейную комбинацию случайных величин

$$\eta_{1\cdot 23\dots n} = \frac{1}{\Lambda_{11}} \sum_{k=1}^{n} \Lambda_{1k}\xi_k. \tag{9}$$

Отсюда, с учетом того, что $E\xi_i = 0$, получаем

$$E\eta_{1\cdot 23\dots n} = 0. \tag{10}$$

$$E\xi_i\eta_{1\cdot 23\ldots n} = \frac{1}{\Lambda_{11}}\sum_{k=1}^{n} E\xi_i\xi_k\Lambda_{1k} = \frac{1}{\Lambda_{11}}\sum_{k=1}^{n}\lambda_{ik}\Lambda_{1k} = \begin{cases} \frac{\Lambda}{\Lambda_{11}}, & i=1, \\ 0, & i\neq 1, \end{cases} \qquad (11)$$

Так как $\sum\limits_{k=1}^{n}\lambda_{1k}\Lambda_{1k} = \Lambda$, а $\sum\limits_{k=1}^{n}\lambda_{ik}\Lambda_{1k} = 0$ равен определителю матрицы, у которой вместо первой стоит i -ая строчка, а если в матрице две строчки одинаковы, ее определитель равен нулю.

Вычислим дисперсию остатка $\sigma^2_{1\cdot 23\ldots n}$, которую будем называть остаточной. С учетом (10) равенства нулю математических ожиданий компонент

$$\sigma^2_{1\cdot 23\ldots n} = E(\eta^2_{1\cdot 23\ldots n}).$$

Заменим одну из η на (9). В силу (11) получаем

$$\sigma^2_{1\cdot 23\ldots n} = E\left(\frac{1}{\Lambda_{11}}\sum_{i=1}^{n}\Lambda_{1i}\xi_i\eta_{1\cdot 23\ldots n}\right) = E(\xi_1\eta_{1\cdot 23\ldots n}) = \frac{\Lambda}{\Lambda_{11}}. \qquad (12)$$

Остаточная дисперсия используется в качетсве меры приближения случайной величины ξ_1 линейной комбинацией $\xi_1^* = \beta_{12}\xi_2 + \ldots + \beta_{1n}\xi_n$.

1.3. Частная корреляция

Частным коэффициентом корреляции между двумя случайными величинами ξ_1 и ξ_2 называется коэффициент корреляции между остатками

$$\varrho_{12\cdot 34\ldots n} = cor(\eta_{1\cdot 34\ldots n}\eta_{2\cdot 34\ldots n}) \qquad (13)$$

имеющий смысл меры линейной зависимости между величинами после устранения изменений, вызванных влиянием $\xi_3, \ldots, \xi_n$. Так как $E\eta_{1\cdot 34\ldots n} = 0$ и $E\eta_{2\cdot 34\ldots n} = 0$, то

$$\varrho_{12\cdot 34\ldots n} = \frac{E(\eta_{1\cdot 34\ldots n}\eta_{2\cdot 34\ldots n})}{\sqrt{E(\eta_{1\cdot 34\ldots n})^2 E(\eta_{2\cdot 34\ldots n})^2}}$$

Итак, в силу (11)

$$E\eta^2_{1\cdot 34\ldots n} = E(\xi_1\eta_{1\cdot 34\ldots n}) = \frac{\Lambda_{22}}{\Lambda_{22\cdot 11}} = \frac{\Lambda_{22}}{\Lambda_{11\cdot 22}},$$
$$E\eta^2_{2\cdot 34\ldots n} = E(\xi_2\eta_{2\cdot 34\ldots n}) = \frac{\Lambda_{11}}{\Lambda_{11\cdot 22}}.$$

Далее

$$E(\eta_{1\cdot 34\ldots n}\eta_{2\cdot 34\ldots n}) = E(\xi_1\eta_{2\cdot 34\ldots n}), \tag{14}$$

так как $\eta_{1\cdot 34\ldots n} = \frac{1}{\Lambda_{22\cdot 11}} \sum\limits_{k=1,k\neq 2}^{n} \Lambda_{22\cdot 1k}\xi_k$ и

$$E\xi_k\eta_{2\cdot 34\ldots n} \begin{cases} = 0, & k = 3, 4, \ldots, n \\ \neq 0, & k = 2 \\ \neq 0, & k = 1 \end{cases}$$

Так как $\eta_{2\cdot 34\ldots n} = \frac{1}{\Lambda_{11\cdot 22}} \sum\limits_{k=2}^{n} \Lambda_{11\cdot 2k}\xi_k$, то (14) имеет вид

$$E(\xi_1\eta_{2\cdot 34\ldots n}) = \frac{1}{\Lambda_{11\cdot 22}} \sum_{k=2}^{n} \Lambda_{11\cdot 2k} E\xi_1\xi_k = \frac{1}{\Lambda_{11\cdot 22}} \sum_{k=2}^{n} \Lambda_{11\cdot 2k}\lambda_{1k} = -\frac{\Lambda_{12}}{\Lambda_{11\cdot 22}}.$$

Таким образом,

$$\varrho_{12\cdot 34\ldots n} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} \tag{15}$$

Между частной корреляцией и соответствующим частным коэффициентом регрессии имеет место линейная зависимость. Согласно (15) и (6) имеем

$$\varrho_{12\cdot 34\ldots n} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} = -\frac{\Lambda_{12}}{\Lambda_{11}}\sqrt{\frac{\Lambda_{11}}{\Lambda_{22}}} = \beta_{12\cdot 34\ldots n}\sqrt{\frac{\Lambda_{11}}{\Lambda_{22}}}. \tag{16}$$

Для случая $n = 3$ матрица вторых моментов Λ имеет вид

$$\Lambda = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\varrho_{12} & \sigma_1\sigma_3\varrho_{13} \\ \sigma_1\sigma_2\varrho_{12} & \sigma_2^2 & \sigma_2\sigma_3\varrho_{23} \\ \sigma_1\sigma_3\varrho_{13} & \sigma_2\sigma_3\varrho_{23} & \sigma_3^2 \end{bmatrix}$$

$$\varrho_{12\cdot 3} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} = \frac{\sigma_1\sigma_2\varrho_{12}\sigma_3^2 - \sigma_1\sigma_3\varrho_{13}\sigma_2\sigma_3\varrho_{23}}{\sigma_2\sigma_3\sqrt{1-\varrho_{23}^2}\sigma_1\sigma_3\sqrt{1-\varrho_{13}^2}} = \frac{\varrho_{12} - \varrho_{13}\varrho_{23}}{\sqrt{1-\varrho_{23}^2}\sqrt{1-\varrho_{13}^2}}$$

ПРИМЕР. В качестве "зависимой"переменной ξ_1 будем рассматривать среднюю продолжительность в году, а в качестве независимых переменных – национальное богатство ξ_2 и объем ξ_3 продаваемой водки за год, приходящийся на одного человека (в бутылках). В качестве индивидов $N = 7$ будем рассматривать годы: 1970, 1975, 1980, 1985, 1990, 1995 и 1998. Звездочками отмечены артефакты или неполные наблюдения, замененные средними.

j	Годы	Продолж-ть жизни ξ_1	Нац.богатсво (млрд.дол.) ξ_2	Водка (бут.в год) ξ_3
1	1970	68.9	2372*	25.3
2	1975	68.1	2372*	28
3	1980	67.6	2489	30
4	1985	69.2	3379	23.5
5	1990	69.2	4130	18
6	1995	64.6	1171	38.4
7	1998	67	689	29.6

Выборочные средние и стандартные оклонения $m_i \pm S_i$ равны соответственно 67.8 ± 1.6, 2372 ± 1182.7, 27.5 ± 6.3. Матрица вторых моментов (ее детерминант равен $\Lambda = 3882374$) и ее алгебраических дополнений имеют вид:

λ_{ij}	ξ_1	ξ_2	ξ_3
ξ_1	2.31	1291	-8.32
ξ_2		1199027	-5294.52
ξ_3			34.40

Λ_{ij}	ξ_1	ξ_2	ξ_3
ξ_1	13214148	-368	138823
ξ_2		10,1	1468
ξ_3			1097860

Коэффициенты корреляции равны $\varrho_{12} = 0.78$, $\varrho_{13} = -0.93$, $\varrho_{23} = -0.82$. Вычисляем частные коэффициены регрессии частные коэффициенты корреляции и дисперсию остатка:

$$\beta_{12} = -\frac{\Lambda_{12}}{\Lambda_{11}} = -\frac{-368}{13214148} = 2.8 \cdot 10^{-5}, \quad \varrho_{12\cdot 3} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} = 0.032,$$

$$\beta_{13} = -\frac{\Lambda_{13}}{\Lambda_{11}} = -\frac{138823}{13214148} = -0.24, \quad \varrho_{13\cdot 2} = -\frac{\Lambda_{13}}{\sqrt{\Lambda_{11}\Lambda_{33}}} = -0.82,$$

$$\sigma^2_{1\cdot 23} = \frac{\Lambda}{\Lambda_{11}} = \frac{3882374}{13214148} = 0.29.$$

Свободный член вычисляется непосредственно $\beta_{11} = m_1 - \beta_{12}m_2 - \beta_{13}m_3 = 74.28$. Наблюдаемая положительная корреляция между продолжительностью жизни и ростом национального богатсва объясняется тем, что при снижении его уровня увеличивается продажа водки, которая и приводит к снижению продолжительности жизни.

1.4. Множественный коэффициент корреляции

Множественным коэффициентом корреляции $\varrho_{1(23\ldots n)}$ между случайной величиной ξ_1 и множеством случайных величин $\xi_2, \ldots, \xi_n$ называется коэффициент корреляции ξ_1 и ее наилучшей среднеквадратической оценкой, построенной по набору случайных величин $\xi_2, \ldots, \xi_n$. В случае $E\xi_i = 0$ множественный коэффициент корреляции имеет вид

$$\varrho_{1(23\ldots n)} = \frac{E(\xi_1 \xi_1^*)}{\sqrt{E\xi_1^2 E(\xi_1^*)^2}}, \tag{17}$$

где $\xi_1^* = \beta_{12}\xi_2 + \ldots + \beta_{1n}\xi_n$ – наилучшая оценка ξ_1 по $\xi_2, \ldots, \xi_n$. По (8) $\xi_1^* = \xi_1 - \eta_{1\cdot 23\ldots n}$, поэтому с учетом (11) и (12)

$$E(\xi_1\xi_1^*) = E(\xi_1(\xi_1 - \eta_{1\cdot 23\ldots n})) = E\xi_1^2 - E\xi_1\eta_{1\cdot 23\ldots n} = \lambda_{11} - \frac{\Lambda}{\Lambda_{11}}$$

$$E(\xi_1^*)^2 = E(\xi_1^2 - 2\xi_1\eta_{1\cdot 23\ldots n} + \eta_{1\cdot 23\ldots n}^2) = \lambda_{11} - \frac{\Lambda}{\Lambda_{11}}.$$

Следовательно,

$$\varrho_{1(23\ldots n)} = \frac{\sqrt{\lambda_{11} - \frac{\Lambda}{\Lambda_{11}}}}{\sqrt{\lambda_{11}}} = \sqrt{1 - \frac{\Lambda}{\Lambda_{11}\lambda_{11}}}. \tag{18}$$

Для численных вычислений оказывается полезной формула

$$\varrho_{1(23\ldots n)}^2 = 1 - \frac{\sigma_{1\cdot 23\ldots n}^2}{\sigma_1^2} = \frac{\sigma_1^2 - \sigma_{1\cdot 23\ldots n}^2}{\sigma_1^2}, \tag{19}$$

так как по (12) $\sigma_{1\cdot 23\ldots n}^2 = \frac{\Lambda}{\Lambda_{11}}$, а $\lambda_{11} = \sigma_1^2$.

Квадрат множественного коэффициента корреляции или *коэффициент детерминации* равен доле дисперсии, объясняемой регрессией, то есть влиянием случайных величин $\xi_2, \ldots, \xi_n$ на ξ_1.

ПРИМЕР (продолжение). Коэффициент детерминации равен

$$R^2 = \varrho_{1(23)}^2 = \frac{\sigma_1^2 - \sigma_{1\cdot 23}^2}{\sigma_1^2} = \frac{2.31 - 0.29}{2.31} = 0.87,$$

множественный коэффициент корреляции равен $\varrho_{1(23)} = 0.93$.

1.5. Отношение дисперсий и частный коэффициент корреляции

$$\frac{\beta_{12}^2\sigma_{2\cdot34\ldots n}^2}{\sigma_{1\cdot234\ldots n}^2} = \frac{\varrho_{12\cdot34\ldots n}^2}{1-\varrho_{12\cdot34\ldots n}^2} \tag{20}$$

Доказательство этого утверждения опирается на равенство Якоби:

$$\Lambda\cdot\Lambda_{11\cdot22} = \Lambda_{11}\cdot\Lambda_{22} - \Lambda_{12}\cdot\Lambda_{21} \tag{21}$$

Используя из (6) и (12) выражения для

$$\beta_{12}^2 = \frac{\Lambda_{12}^2}{\Lambda_{11}^2},$$
$$\sigma_{1\cdot23\ldots n}^2 = \frac{\Lambda}{\Lambda_{11}}, \qquad \sigma_{2\cdot34\ldots n}^2 = \frac{\Lambda_{11}}{\Lambda_{11\cdot22}},$$

получаем левую часть в виде:

$$\frac{\beta_{12}^2\sigma_{2\cdot34\ldots n}^2}{\sigma_{1\cdot234\ldots n}^2} = \frac{\Lambda_{12}^2}{\Lambda_{11}^2}\cdot\frac{\Lambda_{11}}{\Lambda_{11\cdot22}}\cdot\frac{\Lambda_{11}}{\Lambda} = \frac{\Lambda_{12}^2}{\Lambda\cdot\Lambda_{11\cdot22}} = \frac{\Lambda_{12}^2}{\Lambda_{11}\cdot\Lambda_{22}-\Lambda_{12}\cdot\Lambda_{21}}.$$

С другой стороны, из (15) $\varrho_{12\cdot34\ldots n}^2 = \frac{\Lambda_{12}^2}{\Lambda_{11}\Lambda_{22}}$, следовательно,

$$\frac{\varrho_{12\cdot34\ldots n}^2}{1-\varrho_{12\cdot34\ldots n}^2} = \frac{\Lambda_{12}^2}{\Lambda_{11}\Lambda_{22}\left(1-\frac{\Lambda_{12}^2}{\Lambda_{11}\Lambda_{22}}\right)}.$$

Равенство Якоби получается из произведения определителей $(\Lambda_{11}\Lambda_{22} - \Lambda_{12}\Lambda_{21})|\Lambda| = |\Lambda|^2\Lambda_{11\cdot22}$ соответствующих матриц

$$\left[\begin{array}{cc|ccc} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} & \ldots & \Lambda_{1n} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} & \ldots & \Lambda_{2n} \\ \hline \multicolumn{2}{c|}{\mathbb{O}} & \multicolumn{3}{c}{\mathbb{I}} \end{array}\right] \left[\begin{array}{cccc} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \lambda_{n1} & \lambda_{n2} & \ldots & \lambda_{nn} \end{array}\right] = \left[\begin{array}{cc|ccc} |\Lambda| & 0 & 0 & \ldots & 0 \\ 0 & |\Lambda| & 0 & \ldots & 0 \\ \hline \lambda_{31} & \lambda_{32} & \lambda_{33} & \ldots & \lambda_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \lambda_{n1} & \lambda_{n2} & \lambda_{3n} & \ldots & \lambda_{nn} \end{array}\right]$$

2. Проверка гипотез в множественной регрессии

2.1. Оценивание параметров

Решается задача предсказания переменной y с помощью p переменных $x_1, \ldots, x_p$. В пакетах анализа обычно переменную y называют зависимой переменной, а переменные $x_1, \ldots, x_p$ – независимыми. Заметим, что этот термин не эквивалентен независимости в вероятностной смысле. Уравнение модели, выражающее зависимость между y и $x_1, \ldots, x_p$ можно записать в виде

$$y = f(x_1, \ldots, x_p; \beta_1, \ldots, \beta_m) + e,$$

где $\beta_1, \ldots, \beta_m$ – неизвестные параметры, e – ошибка аппроксимации. В частности, если $m = p + 1$

$$f(x_1, \ldots, x_p; \beta_1, \ldots, \beta_m) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \tag{22}$$

то мы имеем модель множественной линейной регрессии. К этой модели сводятся многие другие варианты моделей, если независимые переменные являются функциями, например, $y = \beta_0 + \beta_1 \sin z_1 + \beta_2 \sin z_2 + e$ является моделью множественной линейной регрессии с $x_i = \sin z_i$, или $x_i = x^i$, в этом случае к модели множественной линейной регрессии сводится полиномиальная модель.

Параметры оцениваются по выборке объема n

y	x_1	$\ldots$	x_p
y_1	x_{11}	$\ldots$	x_{p1}
y_2	x_{12}	$\ldots$	x_{p2}
$\vdots$	$\vdots$	$\ldots$	$\vdots$
y_n	x_{1n}	$\ldots$	x_{pn}

Переменные $x_1, \ldots, x_p$ считаются детерминированными. Представим модель множественной линейной регрессии в виде системы уравнений

$$y_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \ldots + \beta_p(x_{pi} - \bar{x}_p) + \delta_i, \quad i = 1, \ldots, n, \tag{23}$$

где независимые $\delta_i \sim \mathcal{N}(0, \sigma)$. Оценки $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p$, которые минимизируют

$$S = \sum_{i=1}^{n}(y_i - \alpha - \beta_1(x_{1i} - \bar{x}_1) - \ldots - \beta_p(x_{pi} - \bar{x}_p))^2, \tag{24}$$

называются частными коэффициентами регрессии. Эта модель сводится к модели (22) выражением для параметра $\beta_0 = \alpha - \beta_1\bar{x}_1 - \ldots - \beta_p\bar{x}_p$. Обозначим через

$$l_{ij} = \frac{1}{n}\sum_{\nu=1}^{n}(x_{i\nu} - \bar{x}_i)(x_{j\nu} - \bar{x}_j), \tag{25}$$

$$l_{0j} = \frac{1}{n}\sum_{\nu=1}^{n}(y_\nu - \bar{y})(x_{j\nu} - \bar{x}_j) \tag{26}$$

оценки вторых центральных моментов, а через L и L_{ij} будем обозначать определитель матрицы вторых выборочных центральных моментов

$$\begin{bmatrix} l_{11} & \ldots & l_{1p} \\ \ldots & \ldots & \ldots \\ l_{p1} & \ldots & l_{pp} \end{bmatrix}$$

и его алгебраические дополнения. Будем предполагать, что $L \neq 0$.

По методу наименьших квадратов оценки параметров имеют вид

$$\hat{\alpha} = \bar{y}, \qquad \hat{\beta}_i = \frac{1}{L}\sum_{j=1}^{p} l_{0j}L_{ij}, \tag{27}$$

а оценкой максимального правдоподобия дисперсии является выражение:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{\nu=1}^{n}(y_\nu - \hat{y}_\nu)^2, \tag{28}$$

где через $\hat{y}_\nu = \hat{\alpha} - \hat{\beta}_1(x_{1\nu} - \bar{x}_1) - \ldots - \hat{\beta}_p(x_{p\nu} - \bar{x}_p)$ обозначено наилучшее линейное предсказание переменной y.

2.2. Моменты оценок частных коэффициентов регрессии

1. $E\hat{\alpha} = \alpha$. Покажем, что

$$\hat{\alpha} = \alpha + \bar{\delta}, \tag{29}$$

В самом деле,

$$\hat{\alpha} = \bar{y} = \frac{1}{n}\sum_{\nu=1}^{n}\left(\alpha + \sum_{i=1}^{p}\beta_i(x_{i\nu} - \bar{x}_i) + \delta_\nu\right) = \alpha + \bar{\delta},$$
$$E\hat{\alpha} = E\left(\alpha + \bar{\delta}\right) = \alpha.$$

2. $E\hat{\beta}_i = \beta_i$. Действительно, из $\bar{y} = \alpha + \bar{\delta}$ получаем, что

$$y_\nu - \bar{y} = \alpha + \sum_{k=1}^{p}\beta_k(x_{k\nu} - \bar{x}_k) + \delta_\nu - \alpha - \bar{\delta} =$$
$$= \sum_{k=1}^{p}\beta_k(x_{k\nu} - \bar{x}_k) + \delta_\nu - \bar{\delta}. \tag{30}$$

Вычисляем математическое ожидание:

$$E\hat{\beta}_i = E\frac{1}{L}\sum_{j=1}^{p} l_{0j}L_{ij} = E\frac{1}{L}\sum_{j=1}^{p} L_{ij}\frac{1}{n}\sum_{\nu=1}^{n}(y_\nu - \bar{y})(x_{j\nu} - \bar{x}_j) =$$
$$= E\frac{1}{L}\sum_{j=1}^{p} L_{ij}\frac{1}{n}\sum_{\nu=1}^{n}(x_{j\nu} - \bar{x}_j)\left(\sum_{k=1}^{p}\beta_k(x_{k\nu} - \bar{x}_k) + \delta_\nu - \bar{\delta}\right) =$$
$$= \frac{1}{L}\sum_{j=1}^{p} L_{ij}\sum_{k=1}^{p}\beta_k\frac{1}{n}\sum_{\nu=1}^{n}(x_{j\nu} - \bar{x}_j)(x_{k\nu} - \bar{x}_k) =$$
$$= \frac{1}{L}\sum_{j=1}^{p} L_{ij}\sum_{k=1}^{p}\beta_k l_{kj} = \frac{1}{L}\sum_{k=1}^{p}\beta_k\sum_{j=1}^{p} L_{ij}l_{kj} = \beta_i,$$

$$\text{так как } \sum_{j=1}^{p} L_{ij}l_{kj} = \begin{cases} 0, & k \neq i, \\ L, & k = i \end{cases}$$

3. $E(\hat{\alpha} - \alpha)^2 = \frac{\sigma^2}{n}$ (очевидно).

4. $E(\hat{\alpha}-\alpha)(\hat{\beta}_i-\beta_i)=0$.

   С одной стороны, $\hat{\alpha}-\alpha=\bar{\delta}$, с другой, из (30) получаем

$$y_\nu-\bar{y}-E(y_\nu-\bar{y})=\delta_\nu-\bar{\delta},$$

$$l_{0j}-El_{0j}=\frac{1}{n}\sum_{\nu=1}^{n}(x_{j\nu}-\bar{x}_j)(y_\nu-\bar{y}-E(y_\nu-\bar{y}))=$$

$$=\frac{1}{n}\sum_{\nu=1}^{n}(x_{j\nu}-\bar{x}_j)(\delta_\nu-\bar{\delta})=\frac{1}{n}\sum_{\nu=1}^{n}(x_{j\nu}-\bar{x}_j)\delta_\nu.$$

   Таким образом,

$$\hat{\beta}_i-\beta_i=\frac{1}{L}\sum_{j=1}^{p}L_{ij}(l_{0j}-El_{0j})=\frac{1}{L}\sum_{j=1}^{p}L_{ij}\frac{1}{n}\sum_{\nu=1}^{n}(x_{j\nu}-\bar{x}_j)\delta_\nu, \quad (31)$$

   следовательно,

$$E(\hat{\alpha}-\alpha)(\hat{\beta}_i-\beta_i)=\frac{1}{Ln}\sum_{j=1}^{p}L_{ij}\sum_{\nu=1}^{n}(x_{j\nu}-\bar{x}_j)E(\bar{\delta}\delta_\nu)=$$

$$=\frac{\sigma^2}{Ln^2}\sum_{j=1}^{p}L_{ij}\sum_{\nu=1}^{n}(x_{j\nu}-\bar{x}_j)=0.$$

5. $E(\hat{\beta}_1-\beta_1)(\hat{\beta}_2-\beta_2)=\frac{\sigma^2L_{12}}{nL}$. Действительно,

$$E\frac{1}{L}\sum_{j=1}^{p}L_{1j}\frac{1}{n}\sum_{\mu=1}^{n}(x_{j\mu}-\bar{x}_j)\delta_\mu\cdot\frac{1}{L}\sum_{i=1}^{p}L_{2i}\frac{1}{n}\sum_{\nu=1}^{n}(x_{i\nu}-\bar{x}_i)\delta_\nu=$$

$$=\frac{\sigma^2}{n^2L^2}\sum_{\mu=1}^{n}\sum_{j=1}^{p}L_{1j}(x_{j\mu}-\bar{x}_j)\sum_{i=1}^{p}L_{2i}(x_{i\mu}-\bar{x}_i)=$$

$$=\frac{\sigma^2}{n^2L^2}\sum_{j=1}^{p}L_{1j}\sum_{i=1}^{p}L_{2i}\sum_{\mu=1}^{n}(x_{j\mu}-\bar{x}_j)(x_{i\mu}-\bar{x}_i)=$$

$$=\frac{\sigma^2}{n^2L^2}\sum_{j=1}^{p}L_{1j}\sum_{i=1}^{p}L_{2i}nl_{ij}=\frac{\sigma^2L_{12}}{nL}.$$

   В частности,

$$D(\hat{\beta}_1)=\frac{\sigma^2L_{11}}{nL}=\frac{\sigma^2}{ns^2_{1\cdot234\ldots p}}. \quad (32)$$

2.3. Соотношения ортогональности для статистик

Оценкой максимального правдоподобия для σ^2 является выборочное значение остаточной дисперсии величины y относительно $x_1, \ldots, x_p$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{\nu=1}^{n}\left(y_\nu - \hat{\alpha} - \sum_{k=1}^{p}\hat{\beta}_k(x_{k\mu} - \bar{x}_k)\right)^2 = s^2_{0\cdot 12\ldots p}. \tag{33}$$

Покажем, что $\hat{\sigma}^2$ независима от $\hat{\alpha}$ и $\hat{\beta}_i$, и что $\frac{n\hat{\sigma}^2}{\sigma^2}$ имеет распределение хи-квадрат с $n-p-1$ степенями свободы.

Рассмотрим сначала частный случай, когда матрица вторых моментов L диагональна.

$$l_{ij} = 0, \quad L_{ij} = 0 \quad \text{при } i \neq j. \tag{34}$$

т.е. $s_i^2 = l_{ii} = \frac{L}{L_{ii}}$, $D(\beta_i) = \frac{\sigma^2}{ns_i^2}$. Выражения $\sqrt{n}(\hat{\alpha} - \alpha)$ и $s_i\sqrt{n}(\hat{\beta}_i - \beta_i)$ есть линейные комбинации от δ_ν. Действительно, согласно (29) и (31),

$$\sqrt{n}(\hat{\alpha} - \alpha) = \frac{1}{\sqrt{n}}\sum_{\nu=1}^{n}\delta_\nu \text{ и}$$

$$s_i\sqrt{n}(\hat{\beta}_i - \beta_i) = s_i\sqrt{n}\left(\frac{1}{L}\sum_{j=1}^{p}L_{ij}\frac{1}{n}\sum_{\mu=1}^{n}(x_{j\mu} - \bar{x}_j)\delta_\mu\right) =$$

$$= s_i\sqrt{n}\left(\frac{L_{ii}}{Ln}\sum_{\mu=1}^{n}(x_{i\mu} - \bar{x}_i)\delta_\mu\right) = \frac{1}{s_i\sqrt{n}}\sum_{\mu=1}^{n}(x_{i\mu} - \bar{x}_i)\delta_\mu$$

Эти линейные комбинации от δ_ν удовлетворяют условиям ортогональности:

$$\sum_{\mu=1}^{n}\left(\frac{1}{\sqrt{n}}\right)^2 = 1,$$

$$\sum_{\mu=1}^{n}\left(\frac{x_{i\mu} - \bar{x}_i}{s_i\sqrt{n}}\right)^2 = 1,$$

$$\sum_{\mu=1}^{n}\left(\frac{x_{i\mu} - \bar{x}_i}{s_i\sqrt{n}\sqrt{n}}\right) = 0.$$

2.4. Разложение суммы квадратов ошибок

Разложим сумму квадратов, объясняемую общей вариацией, на слагаемые, объясняемые вариацией, обусловленной регресией, и вариацией, объясняемой ошибкой:

$$\sum_{\mu=1}^{n}\delta_\mu^2=\sum_{\mu=1}^{n}\left(y_\mu-\alpha-\sum_{k=1}^{p}\beta_k(x_{k\mu}-\bar{x}_k)\right)^2=$$

$$=\sum_{\mu=1}^{n}\left(y_\mu-\alpha-\sum_{k=1}^{p}\beta_k(x_{k\mu}-\bar{x}_k)+\right.$$

$$\left.+\hat{\alpha}-\hat{\alpha}+\sum_{k=1}^{p}\hat{\beta}_k(x_{k\mu}-\bar{x}_k)-\sum_{k=1}^{p}\hat{\beta}_k(x_{k\mu}-\bar{x}_k)\right)^2=$$

$$=\sum_{\mu=1}^{n}\left((y_\mu-\hat{\alpha}-\sum_{k=1}^{p}\hat{\beta}_k(x_{k\mu}-\bar{x}_k))+\right.$$

$$\left.+(\hat{\alpha}-\alpha)+\sum_{k=1}^{p}(\hat{\beta}_k-\beta_k)(x_{k\mu}-\bar{x}_k)\right)^2=$$

$$=n\hat{\sigma}^2+n(\hat{\alpha}-\alpha)^2+n\sum_{k=1}^{p}\sum_{t=1}^{p}l_{kt}(\hat{\beta}_k-\beta_k)(\hat{\beta}_t-\beta_t). \qquad (35)$$

Перекрестные суммы равны нулю:

$$\sum_{\mu=1}^{n}\left(y_\mu-\hat{\alpha}-\sum_{k=1}^{p}\hat{\beta}_k(x_{k\mu}-\bar{x}_k)\right)\left(\sum_{t=1}^{p}(\hat{\beta}_t-\beta_t)(x_{t\mu}-\bar{x}_t)\right)=$$

$$=\sum_{t=1}^{p}(\hat{\beta}_t-\beta_t)\left(\sum_{\mu=1}^{n}(y_\mu-\bar{y})(x_{t\mu}-\bar{x}_t)-\sum_{k=1}^{p}\hat{\beta}_k\sum_{\mu=1}^{n}(x_{k\mu}-\bar{x}_k)(x_{t\mu}-\bar{x}_t)\right)$$

$$\sum_{t=1}^{p}(\hat{\beta}_t-\beta_t)n\left(l_{0t}-\sum_{k=1}^{p}\hat{\beta}_k l_{kt}\right)=0,\ \text{так как}$$

$$\sum_{k=1}^{p}\hat{\beta}_k l_{kt}=\sum_{k=1}^{p}l_{kt}\frac{1}{L}\sum_{j=1}^{p}L_{kj}l_{0j}=\frac{1}{L}\sum_{j=1}^{p}l_{0j}\sum_{k=1}^{p}l_{kt}L_{kj}=l_{0t}.$$

$$\sum_{\mu=1}^{n}\left(\sum_{t=1}^{p}(\hat{\beta}_t-\beta_t)(x_{t\mu}-\bar{x}_t)\right)(\hat{\alpha}-\alpha)=0,$$

$$\sum_{\mu=1}^{n}\left(y_\mu-\hat{\alpha}-\sum_{k=1}^{p}\hat{\beta}_k(x_{k\mu}-\bar{x}_k)\right)(\hat{\alpha}-\alpha)=$$
$$=(\hat{\alpha}-\alpha)\left(n\bar{y}-n\hat{\alpha}-\sum_{k=1}^{p}\hat{\beta}_k\sum_{\mu=1}^{n}(x_{k\mu}-\bar{x}_k)\right)=0,$$

При справедливости (34) последнее слагаемое из (35) имеет вид:

$$\sum_{k=1}^{p}ns_i^2(\hat{\beta}_k-\beta_k)^2,\ \text{так как}\ \sum_{\mu=1}^{n}\left(\sum_{k=1}^{p}(\hat{\beta}_k-\beta_k)(x_{k\mu}-\bar{x}_k)\right)^2=$$
$$=\sum_{\mu=1}^{n}\left(\sum_{k=1}^{p}(\hat{\beta}_k-\beta_k)(x_{k\mu}-\bar{x}_k)\right)\left(\sum_{t=1}^{p}(\hat{\beta}_t-\beta_t)(x_{t\mu}-\bar{x}_t)\right)=$$
$$=\sum_{k=1}^{p}\sum_{t=1}^{p}(\hat{\beta}_k-\beta_k)(\hat{\beta}_t-\beta_t)\sum_{\mu=1}^{n}(x_{k\mu}-\bar{x}_k)(x_{t\mu}-\bar{x}_t)=$$
$$=\sum_{k=1}^{p}\sum_{t=1}^{p}(\hat{\beta}_k-\beta_k)(\hat{\beta}_t-\beta_t)nl_{kt}=\sum_{k=1}^{p}ns_k^2(\hat{\beta}_k-\beta_k)^2.$$

Обозначив через $\eta_0=\sqrt{n}(\hat{\alpha}-\alpha),\ \eta_k=\sqrt{n}s_i(\hat{\beta}_k-\beta_k)$, получаем, что

$$\frac{n\hat{\sigma}^2}{\sigma^2}=\frac{1}{\sigma^2}\left(\sum_{\mu=1}^{n}\delta_\mu^2-\eta_0^2-\sum_{k=1}^{p}\eta_k^2\right)\ \sim\ \chi^2(n-p-1). \tag{36}$$

2.5. Проверка значимости коэффициентов регрессии

В общем случае, когда матрица вторых моментов не является диагональной, величины x_i следует при помощи ортогонального преобразования заменить на величины x_i', для которых матрица вторых моментов будет диагональной. К параметрам β_i следует применить контрградиентное преобразование. Т.е. если $Y=A+B^T(X-\bar{X})$ и $(X-\bar{X})=CZ$, где Z – вектор с некоррелированными компонентами, то $Y=A+B^TCZ=A+(C^TB)^TZ$ и $B_1=C^TB$, $B=CB_1$.

Отсюда следует, что величины

$$t_0 = \frac{(\hat{\alpha}-\alpha)\sqrt{n}}{\sigma\sqrt{\frac{1}{n-p-1}\frac{n\hat{\sigma}^2}{\sigma^2}}} = \frac{\hat{\alpha}-\alpha}{s_{0\cdot 12\ldots p}}\sqrt{n-p-1},$$

$$t_1 = \frac{(\hat{\beta}_1-\beta_1)\sqrt{n}s_{1\cdot 23\ldots p}}{\sigma\sqrt{\frac{1}{n-p-1}\frac{n\hat{\sigma}^2}{\sigma^2}}} = \frac{(\hat{\beta}_1-\beta_1)s_{1\cdot 23\ldots p}}{s_{0\cdot 12\ldots p}}\sqrt{n-p-1}, \qquad (37)$$

и аналогичные величины с $\beta_2, \ldots, \beta_p$ имеют распределение Стьюдента с $n-p-1$ степенями свободы. Распределения этих статистик используются для построения доверительных интервалов параметров и проверки гипотез о равенстве нулю частного коэффициента регрессии. Гипотеза $H_0 : \beta_1 = 0$ может быть использована для проверки равенства нулю частного коэффициента корреляции, но вместо статистики (37), согласно (20) используется эквивалентная ей статистика вида

$$\frac{\hat{\varrho}_{01\cdot 234\ldots p}}{\sqrt{1-\varrho^2_{01\cdot 234\ldots p}}}\sqrt{n-p-1} \;\sim\; T(n-p-1). \qquad (38)$$

Обозначим через $Q_2 = \sum\limits_{\nu=1}^{n}\left(y_\nu - \hat{\alpha} - \sum\limits_{k=1}^{p}\hat{\beta}_k(x_{k\mu}-\bar{x}_k)\right)^2$. Если использовать несмещенную оценку дисперсии

$$S^2 = \frac{1}{n-p-1}Q_2,$$

то статистики t_0 и t_1 можно записать в виде

$$t_0 = \frac{\hat{\alpha}-\alpha}{S}\sqrt{n}\,,$$

$$t_1 = \frac{(\hat{\beta}_1-\beta_1)s_{1\cdot 23\ldots p}}{S}\sqrt{n} = \frac{(\hat{\beta}_1-\beta_1)}{\hat{\sigma_{\beta_1}}}\,.$$

Выражение $\frac{S}{\sqrt{n}s_{1\cdot 2,3,\ldots,p}}$ соответствует оценке стандартного отклонения σ_{β_1} оценки $\hat{\beta}_1$ из (32). Поэтому при вычислениях используются дисперсии оценок частных коэффициентов регрессии, на корни из которых делятся оценки параметров для получения нужных статистик Стьюдента.

2.6. Проверка значимости прогноза

Для проверки гипотезы о значимости прогноза $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ используется то, что $\frac{n}{\sigma^2}\sum\limits_{k=1}^{p}\sum\limits_{t=1}^{p} l_{kt}(\hat{\beta}_k - \beta_k)(\hat{\beta}_t - \beta_t)$ имеет распределение хи-квадрат с p степенями свободы. При (34) это очевидно,

$$n\sum_{k=1}^{p}\sum_{t=1}^{p} l_{kt}(\hat{\beta}_k - \beta_k)(\hat{\beta}_t - \beta_t) = \sum_{k=1}^{p} n s_k^2(\hat{\beta}_k - \beta_k)^2,$$

$$\text{где} \quad \sqrt{n}s_k(\hat{\beta}_k - \beta_k) \sim \mathcal{N}(0, \sigma).$$

В общем случае используем утверждение о том, что для любого нормального вектора Z с нулевым средним и ковариационной матрицей $\Sigma = EZZ^T$ справедливо $Z^T\Sigma^{-1}Z \ \sim \ \chi^2(p)$. Действительно, пусть $Y = BZ$ вектор с независимыми компонентами, так что $Y^TY \ \sim \ \chi^2(p)$. С другой стороны,

$$Y^TY = Z^T(B^TB)Z \ \sim \ \chi^2(p),$$

а ковариационная матрица вектора Z с коррелируемыми компонентами имеет тот же вид

$$\Sigma = EZZ^T = EB^{-1}YY^T(B^{-1})^T = B^{-1}(B^{-1})^T = (B^TB)^{-1}.$$

Поскольку ковариационная матрица вектора параметров (п.2.2, 5) имеет вид:

$$cov(\beta) = \left\{\frac{\sigma^2 L_{ij}}{nL}\right\}_{i,j=1}^{p} = \frac{\sigma^2}{n}L^{-1}, \quad \text{то}$$

$$(\hat{\beta} - \beta)^T\left(\frac{\sigma^2}{n}L^{-1}\right)^{-1}(\hat{\beta} - \beta) = \frac{n}{\sigma^2}(\hat{\beta} - \beta)^T L(\hat{\beta} - \beta) =$$

$$= \frac{n}{\sigma^2}\sum_{k=1}^{p}\sum_{t=1}^{p} l_{kt}(\hat{\beta}_k - \beta_k)(\hat{\beta}_t - \beta_t) \ \sim \ \chi^2(p).$$

Следовательно, статистика

$$F = \frac{n-p-1}{p} \cdot \frac{\sum\limits_{k=1}^{p}\sum\limits_{t=1}^{p} l_{kt}(\hat{\beta}_k - \beta_k)(\hat{\beta}_t - \beta_t)}{s^2_{0\cdot 12\ldots p}} \ \sim \ F(p, n-p-1). \qquad (39)$$

Статистику F можно выразить через множественный коэффициент корреляции. Для этого, используя то, что при $\beta_1 = \ldots = \beta_p = 0$ выражение $l_{00} = \frac{1}{n}\sum\limits_{\nu=1}^{n}(y_\nu - \bar{y})^2$ согласно (30) превращается в

$$l_{00} = \frac{1}{n}\sum_{\nu=1}^{n}(\delta_\nu - \bar{\delta})^2 = \frac{1}{n}\sum_{\nu=1}^{n}\delta_\nu^2 - \bar{\delta}^2, \text{ из (35) получаем}$$

$$\sum_{k=1}^{p}\sum_{t=1}^{p} l_{kt}\hat{\beta}_k\hat{\beta}_t = \frac{1}{n}\sum_{\nu=1}^{n}\delta_\nu^2 - \hat{\sigma}^2 - (\hat{\alpha}-\alpha)^2 =$$

$$= l_{00} - \hat{\sigma}^2 = l_{00}\left(1 - \frac{s^2_{0\cdot 12\ldots p}}{l_{00}}\right) = l_{00}r^2_{0(12\ldots p)}.$$

А так как $s^2_{0\cdot 12\ldots p} = l_{00}(1 - r^2_{0(12\ldots p)})$, то статистику (39) представим в виде:

$$F = \frac{n-p-1}{p}\cdot\frac{r^2_{0(12\ldots p)}}{1-r^2_{0(12\ldots p)}} \;\sim\; F(p, n-p-1). \tag{40}$$

Таблица анализа дисперсий

Обозначим через $\hat{\mathbf{Y}} = Xb$ вектор наилучшего в среднеквадратичном смысле линейного предсказания переменной Y по переменным $X_1, \ldots, X_p$.

Источник дисперсии	Сумма квадратов	степени свободы	средний квадрат
Регрессия	$SS_R = \sum\limits_{i=1}^{n}(\hat{\mathbf{y}}_i - \bar{y})^2$	$\nu_R = p$	$MS_R = SS_R/\nu_R$
Откл.от регр.	$SS_E = \sum\limits_{i=1}^{n}(y_i - \hat{\mathbf{y}}_i)^2$	$\nu_E = n-p-1$	$MS_E = SS_E/\nu_E$
полная	$SS_T = \sum\limits_{i=1}^{n}(y_i - \bar{y})^2$	$\nu_T = n-1$	$MS_T = SS_T/\nu_T$

Таблица 1. Дисперсионный анализ для модели множественной регрессии ($SS_E = SS_T - SS_R$)

Из таблицы 1 дисперсионного анализа для модели множественной регрессии может быть получена оценка коэффициента детерминации $R^2 =$

$\frac{SS_R}{SS_T}$ и выражение для статистики $F = \frac{MS_R}{MS_E}$, которая при справедливости нулевой гипотезы о равенстве нулю частных коэффициентов регрессии имеет распределение Фишера с p и $n - p - 1$ степенями свободы.

2.7. Задача регрессии в матричном виде

Уравнения линейной регрессии представим в виде $Y = X\beta + e$, где зависимая переменная задается вектором $Y = (y_1, \ldots, y_n)^T$, матрица

$$X^T = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_{11} & x_{12} & \ldots & x_{1n} \\ & & \ldots & \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{bmatrix} \tag{41}$$

называется матрицей плана, $\beta = (\beta_0, \ldots, \beta_p)^T$ – вектор параметров, e – нормально распределенный вектор с нулевым средним и ковариационной матрицей вида $\sigma^2\mathbf{I}$.

При минимизации квадратичной формы $S = (Y - X\beta)^T(Y - X\beta)$ вектор оценок $b = (b_0, \ldots, b_p)^T$ является решением системы нормальных уравнений

$$(X^T X)\beta = X^T Y \tag{42}$$

Действительно, дифференцирование по вектору параметров приводит к

$$2X^T(Y - X\beta) = 0 \Longleftrightarrow X^T Y = X^T X\beta.$$

Следовательно,

$$b = (X^T X)^{-1} X^T Y. \tag{43}$$

Оценка параметров

Покажем, что оценки (27) вида

$$b_i = \frac{1}{L}\sum_{j=1}^{p} l_{0j}L_{ij},\ i = 1, 2, \ldots, p,$$

$$b_0 = \bar{y} - \sum_{i=1}^{p} b_i\bar{x}_i$$

удовлетворяют уравнению

$$(X^TX)b = X^TY\,.$$

Действительно,

$$X^TY = \begin{bmatrix} n\bar{y} \\ \sum\limits_{j=1}^{n} x_{1j}y_j \\ \vdots \\ \sum\limits_{j=1}^{n} x_{pj}y_j \end{bmatrix},\quad X^TX = \begin{bmatrix} n & n\bar{x}_1 & \ldots & n\bar{x}_p \\ n\bar{x}_1 & \sum\limits_{j=1}^{n} x_{1j}^2 & \ldots & \sum\limits_{j=1}^{n} x_{1j}x_{pj} \\ \vdots & & & \\ n\bar{x}_p & \sum\limits_{j=1}^{n} x_{pj}x_{1j} & \ldots & \sum\limits_{j=1}^{n} x_{pj}^2 \end{bmatrix} \tag{44}$$

Первый элемент вектора X^TXb имеет вид

$$n(\bar{y} - \sum_{i=1}^{p} b_i\bar{x}_i) + \sum_{i=1}^{p} b_i n\bar{x}_i = n\bar{y}\,.$$

Второй элемент равен

$$n\bar{x}_1(\bar{y} - \sum_{i=1}^{p} b_i\bar{x}_i) + \sum_{i=1}^{p} b_i \sum_{j=1}^{n} x_{1j}x_{ij} =$$

$$= n\bar{x}_1\bar{y} + \sum_{i=1}^{p} b_i \left(\sum_{j=1}^{n} x_{1j}x_{ij} - n\bar{x}_1\bar{x}_i \right) = n\bar{x}_1\bar{y} + n\sum_{i=1}^{p} b_i l_{1i}\,.$$

Если вместо b_i подставить оценки и поменять местами суммирование, то

$$\sum_{i=1}^{p} b_i l_{1i} = \sum_{i=1}^{p} l_{1i}\frac{1}{L}\sum_{j=1}^{p} l_{0j}L_{ij} = \frac{1}{L}\sum_{j=1}^{p} l_{0j}\sum_{i=1}^{p} l_{1i}L_{ij} = l_{01}$$

Таким образом второй элемент вектора X^TXb совпадает со вторым элементом вектора X^TY

$$n\bar{x}_1\bar{y} + nl_{01} = n\bar{x}_1\bar{y} + \sum_{j=1}^{n} y_j x_{1j} - n\bar{y}\bar{x}_1 = \sum_{j=1}^{n} y_j x_{1j}\,.$$

Для остальных индексов все аналогично.

Ковариация оценок

Ковариационная матрица оценок частных коэффициентов регрессии можно представить в виде:

$$\begin{aligned} Cov(b) &= \mathbf{E}(b - \mathbf{E}b)(b - \mathbf{E}b)^T = \\ &= \mathbf{E}((X^TX)^{-1}X^T(Y - \mathbf{E}Y))((X^TX)^{-1}X^T(Y - \mathbf{E}Y))^T = \\ &= (X^TX)^{-1}X^TCov(Y)((X^TX)^{-1}X^T)^T = \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}\,. \end{aligned}$$

Учитывая выражения для моментов из п.2.2, построим ковариационную матрицу $cov(b)$. Заметим, что

$$\begin{aligned} \mathbf{D}(b_0) &= \mathbf{D}\left(\hat{\alpha} - \sum_{i=1}^{p} b_i\bar{x}_i\right) = \\ &= \mathbf{D}\hat{\alpha} + \mathbf{D}\left(\sum_{i=1}^{p} b_i\bar{x}_i\right) - 2cov\left(\hat{\alpha}, \sum_{i=1}^{p} b_i\bar{x}_i\right) = \\ &= \frac{\sigma^2}{n} + \sum_{i,j=1}^{p} \bar{x}_i\bar{x}_j\frac{L_{ij}\sigma^2}{nL}\,, \\ cov(b_0, b_j) &= cov\left(\hat{\alpha} - \sum_{i=1}^{p} b_i\bar{x}_i, b_j\right) = -\sum_{i=1}^{p} \bar{x}_i\frac{L_{ij}\sigma^2}{nL}\,. \end{aligned}$$

Соберем ковариационную матрицу.

$$cov(b) = \frac{\sigma^2}{n}\begin{bmatrix} 1+\frac{1}{L}\sum\limits_{i,j=1}^{p}\bar{x}_i\bar{x}_jL_{ij} & -\frac{1}{L}\sum\limits_{i=1}^{p}\bar{x}_iL_{i1} & \dots & -\frac{1}{L}\sum\limits_{i=1}^{p}\bar{x}_iL_{ip} \\ -\frac{1}{L}\sum\limits_{i=1}^{p}\bar{x}_iL_{i1} & \frac{L_{11}}{L} & \dots & \frac{L_{1p}}{L} \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{L}\sum\limits_{i=1}^{p}\bar{x}_iL_{ip} & \frac{L_{p1}}{L} & \dots & \frac{L_{pp}}{L} \end{bmatrix}$$

Покажем, что $(X^TX)cov(b) = \sigma^2\mathbf{I}$. При умножении строки $n(1, \bar{x}_1, \dots, \bar{x}_p)$ матрицы X^TX на первый столбец матрицы $cov(b)$ получаем σ^2, на все остальные столбцы ноль.

Так как $l_{kt} = \frac{1}{n}\sum_{j=1}^{n}x_{kj}x_{tj} - \bar{x}_k\bar{x}_t$, вторая строка вида

$$\left(n\bar{x}_1, \sum_{j=1}^{n}x_{1j}^2, \sum_{j=1}^{n}x_{1j}x_{2j}, \dots, \sum_{j=1}^{n}x_{1j}x_{pj}\right) = \\ = n(\bar{x}_1, l_{11}+\bar{x}_1^2, l_{12}+\bar{x}_1\bar{x}_2, \dots, l_{1p}+\bar{x}_1\bar{x}_p)$$

при умножении на второй столбец матрицы $cov(b)$ дает выражение

$$\sigma^2\left(-\frac{\bar{x}_1}{L}\sum_{i=1}^{p}\bar{x}_iL_{ik} + \sum_{k=1}^{p}(l_{1k}+\bar{x}_1\bar{x}_k)\frac{L_{k1}}{L}\right) = \sigma^2,$$

на первый и на остальные далее второго $m > 2$

$$\sigma^2\left(\bar{x}_1 + \frac{\bar{x}_1}{L}\sum_{i,k=1}^{p}\bar{x}_i\bar{x}_kL_{ik} + \sum_{k=1}^{p}(l_{1k}+\bar{x}_1\bar{x}_k)\left(-\frac{1}{L}\sum_{i=1}^{p}\bar{x}_iL_{ik}\right)\right) = 0,$$

$$\sigma^2\left(-\frac{\bar{x}_1}{L}\sum_{i=1}^{p}\bar{x}_iL_{im} + \sum_{k=1}^{p}(l_{1k}+\bar{x}_1\bar{x}_k)\frac{L_{km}}{L}\right) = 0.$$

Остаточная сумма квадратов

Выясним, насколько согласовано с выражением для остаточной дисперсии $\sigma^2_{1\cdot 23\ldots n} = \frac{\Lambda}{\Lambda_{11}}$ выражение для суммы квадратов остатков

$$\begin{aligned}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) &= Y^TY - Y^TX\hat{\beta} = \\ &= Y^TY - Y^TX(X^TX)^{-1}X^TY = \\ &= \begin{vmatrix} Y^TY & Y^TX \\ X^TY & X^TX \end{vmatrix} / |X^TX| = \frac{1}{G_{11}},\end{aligned} \tag{45}$$

где $G = S^{-1}$, $S = Z^TZ$ матрица вторых выборочных моментов с точностью до множителя n, $Z = [Y, X]$ – матрица размерности n на $p+1$, X – матрица плана размерности n на p. Заметим, что сумма квадратов остатков не зависит от смещений, поэтому можно говорить о центральных моментах. Здесь мы воспользовались разложением определителя матрицы блочной структуры

$$|X| = |D|(a - b^TD^{-1}c) = a|D| - \sum_i b_i \sum_k D_{ki}c_k,$$

$$\text{где} \quad X = \begin{bmatrix} a & b^T \\ c & D \end{bmatrix}.$$

3. Проверка нормальности остатков

Рассмотрим подробнее процедуру построения графика соотношения эмпирических и теоретических квантилей $QQ - plot$. Пусть имеется выборка $x_1, \ldots, x_n$, распределение которой нас интересует. Для примера, рассмотрим данные, полученные в результате моделирования в R при $n = 100$

$$x < -rnorm(n, mean = 1, sd = 2)$$

Построим ее эмпирическую функцию распределения, при помощи функции $plot(ecdf(x))$, график функции распределения, например, $curve(pnorm, -3, 3)$ и отметим эмпирические Qe и теоретические квантили Qt для всех значений $F^{-1}\left(\frac{i-0.5}{n}\right)$, где $F(x)$ функция распределения. Половинка вычитается для того чтобы не получить $F^{-1}(1) = +\infty$.

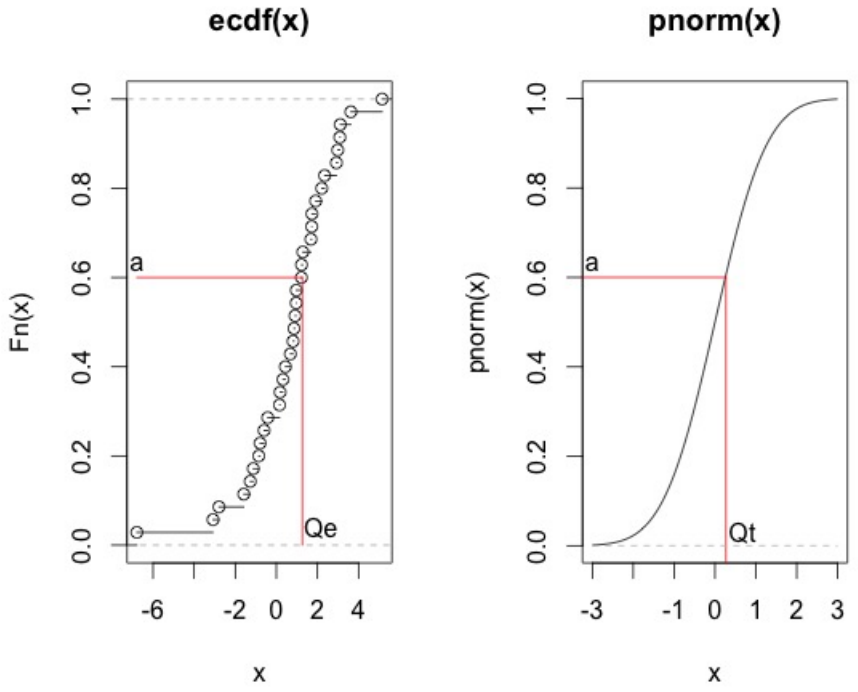
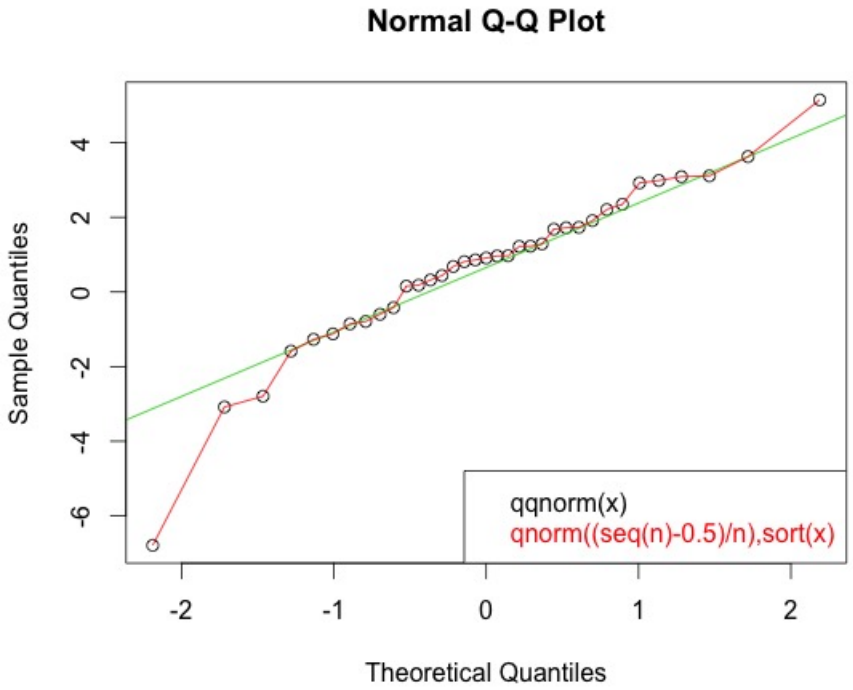
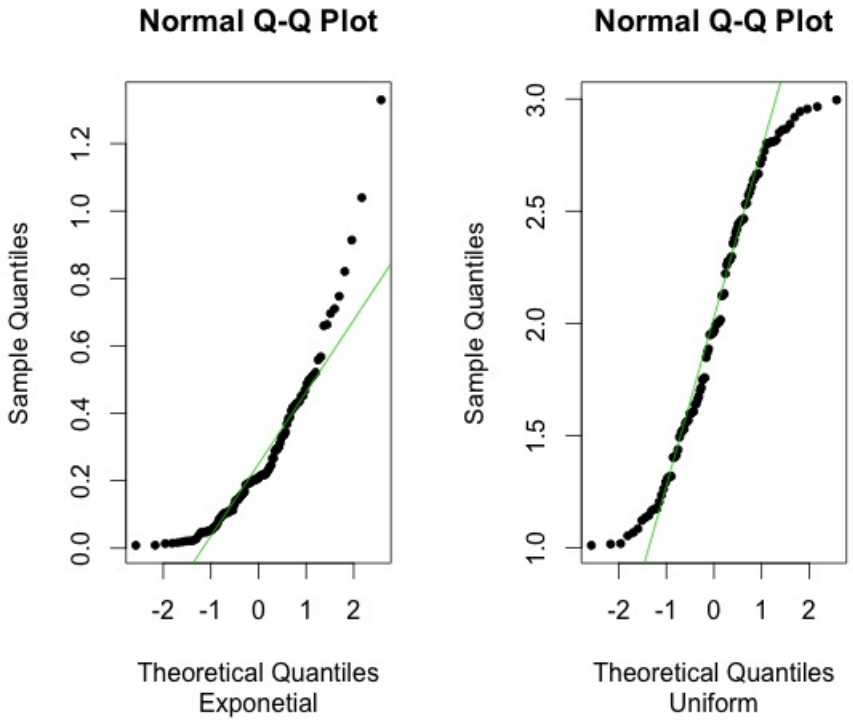


График квантиль-квантиль строится при помощи функции $qqnorm(x)$. По вертикальной оси откладываются упорядоченные значения $x_1, \ldots, x_n$, по горизонтальной $F^{-1}\left(\frac{i-0.5}{n}\right)$, в случае нормального распределения

$$qnorm((seq(n) - 0.5)/n).$$





Примеры графиков квантиль-квантиль с теоретическими нормальными квантилями и эмпирическими квантилями равномерного и экспоненциального распределений представлены на рисунках.

4. Пошаговый регрессионный анализ

Задачей пошаговой регрессии является упорядочивание независимых переменных $X_1, \ldots, X_p$ по важности их для предсказания Y. Это означает, что необходимо найти такое подмножество из независимых переменных, которому бы соответствовал максимальный множественный коэффициент корреляции. Можно просто рассмотреть всевозможные подмножества, но количество необходимых уравнений регрессии существенно увеличивается с возрастанием числа независимых переменных p. Число этих уравнений определяется как сумма всевозможных чисел сочетаний из p:

$$C_p^1 + C_p^2 + \ldots + C_p^p = 2^p - 1.$$

Решение проблемы заключается в пошаговой регрессии, когда переменные включаются в рассмотрение по предварительному критерию. Совокупность критериев, определяющих, какие переменные включать, заменять, удалять, называется пошаговой процедурой.

Стандартная пошаговая процедура заключается в следующем. Пусть имеется набор S_k из k независимых переменных. Для переменных $X_i \notin S_k$ проверяется гипотеза $H_0 : r_{yi \cdot S_k} = 0$ о равенстве нулю частного коэффициента корреляции между переменными Y и X_i при фиксированных значениях переменных из набора S_k. Статистика F-включения по (40) имеет вид

$$F_{yi \cdot S_k} = \frac{r^2_{yi \cdot S_k}(n - k - 2)}{1 - r^2_{yi \cdot S_k}} \tag{46}$$

Гипотеза означает, что *предсказание Y не улучшается при включении X_i*. Если нулевая гипотеза верна, то эта статистика имеет F-распределение с 1 и $n - k - 2$ степенями свободы. Если нулевая гипотеза отвергается в случае нескольких переменных (статистика F-включения больше заранее

заданного значения минимума F-включения), то включается переменная, у который значение статистики F-включения оказывается наибольшим.

Статистика F-удаления применяется для переменных из множества S_k. Для каждой переменной $X_i \in S_k$ проверяется гипотеза $H_0 : r_{yi \cdot S_{k-1}} = 0$, где S_{k-1} есть множество S_k без элемента X_i. Эта гипотеза означает, что *набор S_{k-1} предсказывает Y также хорошо, как и набор S_k.*

$$F_{yi \cdot S_{k-1}} = \frac{r^2_{yi \cdot S_{k-1}}(n - k' - 2)}{1 - r^2_{yi \cdot S_{k-1}}}, \tag{47}$$

где $k' = k - 1$. Если гипотеза верна, то эта статистика имеет $F(1, n - k' - 2)$- распределение. Элемент X_i удаляется, если гипотеза принимается, то есть F - удаления меньше заранее оговоренного максимума, который для успешной работы процедуры принимается меньшим значения минимума F-включения.

ШАГ 0. Вычисляются простые коэффициенты корреляции r_{yi} и величины F-включения F_{yi} для $i = 1, \ldots, p$. Этому соответствует случай $k = 0$ при пустом наборе S_k. Статистика критерия получается из (46) подстановкой $k = 0$. Величина F_{yi} имеет F-распределение с 1 и $n - 2$ степенями свободы и служит для проверки гипотезы $H_0 : \varrho_{yi} = 0,\ i = 1, \ldots, p$.

ШАГ 1. Переменная X_{i_1}, которой отвечает наибольшее значение статистики F-включения (что эквивалентно наибольшему значению квадрата коэффициента корреляции с Y), выбирается как наилучший предиктор для Y. Вычисляются соответствующее уравнение наименьших квадратов, таблица дисперсионного анализа и множественный коэффициент корреляции $R_{y \cdot i_1}$. Далее для проверки гипотезы $H_0 : \varrho_{yi \cdot i_1} = 0$, $i = 1, \ldots, p,\ i \neq i_1$, вычисляются частные коэффициенты корреляции $r_{yi \cdot i_1}$ и значения статистики F-включения

$$F_{yi \cdot i_1} = \frac{r^2_{yi \cdot i_1}(n - 3)}{1 - r^2_{yi \cdot i_1}}, \tag{48}$$

которая имеет 1 и $n-3$ степени свободы. Если F-включения меньше установленного минимума, то выполняется шаг S, иначе шаг 2.

ШАГ 2. Переменная X_{i_2}, которой отвечает наибольшее значение статистики F-включения (что эквивалентно наибольшему значению квадрата коэффициента корреляции с Y), выбирается как наилучший предиктор для Y при условии, что уже выбрана переменная X_{i_1}. Вычисляются соответствующее уравнение наименьших квадратов, таблица дисперсионного анализа и множественный коэффициент корреляции $R_{y\cdot i_1,i_2}$. Далее вычисляются частные коэффициенты корреляции $r_{yi\cdot i_1,i_2}$, значения F-удаления

$$F_{yi_1\cdot i_2}=\frac{r^2_{yi_1\cdot i_2}(n-3)}{1-r^2_{yi_1\cdot i_2}},\quad F_{yi_2\cdot i_1}=\frac{r^2_{yi_2\cdot i_1}(n-3)}{1-r^2_{yi_2\cdot i_1}}. \tag{49}$$

Они используются для проверки гипотезы $H_0:\varrho_{yi_1\cdot i_2}=0$ и $H_0:\varrho_{yi_2\cdot i_1}=0$ соответственно. Наконец, вычисляется частный коэффициент корреляции $\varrho_{yi\cdot i_1i_2}=0$ для проверки гипотезы $H_0:\varrho_{yi\cdot i_1i_2}=0$ и значения F-включения

$$F_{yi\cdot i_1i_2}=\frac{r^2_{yi\cdot i_1i_2}(n-4)}{1-r^2_{yi\cdot i_1i_2}} \tag{50}$$

для всех $i=1,\ldots,p,\ i\neq i_1,\neq i_2$. Если все значения F-включения меньше установленного минимума, то выполняется шаг S иначе шаг 3.

ШАГ 3. А) Пусть L обозначает набор из l независимых переменных, которые включены в уравнение регрессии. Если какое-либо из значений F-удаления для переменных из L меньше установленного минимума, то переменная с наименьшим значением F-удаления исключается из набора и выполняется шаг 3б) с заменой l на $l-1$.

Если для всех переменных, не входящих в набор L, значения F-включения меньше установленного минимума включения, то выполняется шаг S. В противном случае добавляется переменная с максимальным F-включением и l заменяется на $l+1$.

Б) Вычисляются уравнение наименьших квадратов, таблица дисперсионного анализа и множественный коэффициент корреляции $R_{y\cdot l}$. Далее вычисляются частные коэффициенты корреляции $r_{yi_j\cdot(l-1)}$, значения F-удаления $F_{yi_j\cdot(l-1)}$ для переменных из набора L. Проверяются гипотезы равенства нулю частных коэффициентов корреляции.

Наконец, вычисляется частный коэффициент корреляции $\varrho_{yi\cdot l} = 0$ для проверки гипотезы $H_0 : \varrho_{yi\cdot i_1 i_2} = 0$ и значения F-включения

$$F_{yi\cdot l} = \frac{r^2_{yi\cdot l}(n-l-2)}{1-r^2_{yi\cdot l}} \tag{51}$$

для всех $i = 1, \ldots, p,\ i \notin L$. Эта статистика имеет 1 и $n-l-2$ степени свободы. Если все значения F-включения меньше установленного минимума, то выполняется шаг S, в противном случае следующий шаг.

ШАГ S. Для каждого шага печатается номер шага, номера включенных и исключенных переменных, значения F-включения, F-удаления и множественного коэффициента корреляции между Y и включенными переменными.

ПРИМЕР (продолжение).y - продолжительность жизни, x_1 - объем национального богатства, x_2 - количество продаваемых за год бутылок водки, приходящееся на одного человека, уравнение регрессии:

$$y = 74.27634 + 0.00003x_1 - 0.23754x_2,$$

множественный коэффициент корреляции $R = 0.9341$, коэффициент детерминации $R^2 = 0.8726$, $p < 0.01624$ значимость статистики Фишера $F(2,4) = 13.696$, частные коэффициенты корреляции:

$$r_{12\cdot 3} = 0.0318,\ p = 0.9523, r_{13\cdot 2} = -0.824,\ p = 0.0437$$

.

4.1. Задание по теме „Множественная регрессия“

Для того чтобы сделать обработку по множественной регрессии в программе $StatSoft$, нужно запустить программу *Statistics - Multiple Regression*. Далее нужно выбрать переменные ($Variables$). При необходимости привлечения пошаговой процедуры ставится галочка в *Advanced options* (*stepwise or ridge regression*). Далее OK. В таком случае нужно будет выбрать метод: $Forward$ или $Backward$ в зависимости от того, происходит ли отбор информативных или исключение малоинформативных признаков. Коэффициенты регрессии, их значимость можно получить из $Summary$: *Regression results*, таблица дисперсионного анализа из $ANOVA$ ($Overal\ goodnessof fit$), частные коэффициенты регрессии соответственно из *Partial correlations*.

1. ASIAN% – процент лиц азиатского происхождения;
2. BLACK% – процент афроамериканцев;
3. TEMPER – среднеиюльская температура;
4. GROSS – средеяя арендная плата;
5. PRECEP – годовое количество осадков;
6. TRANSP – процент лиц, использующих в городе общественный транспорт;
7. DEGREE – процент людей с образованием бакалавр и выше.

Построить уравнение регрессии по данным из табл.2, взяв в качестве зависимой переменной один признак (номер варианта), в качестве независимых – все остальные. Проверить значимость регрессии и частных коэффициентов, построить частные коэффициенты корреляции. Для проверки вычислить коэффициенты корреляции между остатком и вычитаемыми переменными.

Вариант		1	2	3	4	5	6	7
city	st	asian%	black%	temper	cross	precep	transp	degree
COLUMBUS	1	2,4	22,6	73,2	422	38,1	4,7	24,6
CLEVELAN	1	1	46,6	71,9	322	36,6	14,1	8,1
CINCINNA	1	1,1	37,9	76,4	329	40,7	11,2	22,2
TOLEDO	1	1	19,7	72,1	378	33	3	14,1
AKRON	1	1,2	24,5	71,9	360	36,8	3,3	14,9
HOUSTON	2	4,1	28,1	83,5	390	50,8	6,5	25,1
DALLAS	2	2,2	29,5	85,9	426	36,1	6,7	27,1
SAN ANTO	2	1,1	7	85	369	31	4,9	17,8
EL PASO	2	1,2	3,4	82,3	349	8,8	3,1	16,2
AUSTIN	2	3	12,4	84,5	410	31,9	5,1	34,4
FORT WOR	2	2	22	85,3	403	33,7	1,7	21,5
ARLINGTO	2	3,9	8,4	85,3	444	33,7	0,2	30
CORPUS C	2	0,9	4,8	84,1	373	30,1	1,9	17,8
LOS ANGE	3	9,8	14	74,3	600	14,8	10,5	23
SAN DIEG	3	11,8	9,4	71	602	9,9	4,2	29,8
SAN JOSE	3	19,5	4,7	69,5	755	14,4	3,5	25,3
SAN FRAN	3	29,1	10,9	59,1	653	19,7	33,5	35
LONG BEA	3	13,6	13,7	73,1	605	11,8	6	23,2
SACRAMEN	3	15	15,3	75,7	495	17,5	4	23,5
FRESNO	3	12,5	8,3	81,9	441	10,6	2,1	19,1
OAKLAND	3	14,8	43,9	62,1	538	24,3	17,9	27,2
SANTA AN	3	9,7	2,6	72,6	736	12,3	8,4	10,6
ANAHEIM	3	9,4	2,5	72,6	712	12,3	3,8	18,8
RIVERSID	3	5,2	7,4	77,9	575	9,6	1,6	19,3
STOCKTON	3	22,8	9,6	77,7	476	14	2	15

Таблица 2. Статистические данные об американских городах.

5. Распределение Уишарта и его применение

5.1. Определение центрального распределения Уишарта и его свойства

Пусть векторы $U_j \sim \mathcal{N}_p(0, \Sigma)$, $j = 1, \ldots, n$, независимы.

	Y_1	$\ldots$	Y_p	Y
U_1^T	U_{11}	$\ldots$	U_{p1}	$U_1^T L = L^T U_1$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
U_n^T	U_{1n}	$\ldots$	U_{pn}	$U_n^T L = L^T U_n$

$$\mathcal{U} = \begin{bmatrix} U_{11} & \ldots & U_{p1} \\ \vdots & \ldots & \vdots \\ U_{1n} & \ldots & U_{pn} \end{bmatrix}$$

Векторы $U_j = (U_{1j}, \ldots, U_{pj})^T$, $j = 1, \ldots, n$ соответствуют n индивидам, векторы $Y_i = (U_{i1}, \ldots, U_{in})^T$, $i = 1, \ldots, p$, p признакам. $L = (L_1, \ldots, L_p)$ коэффициенты линейной комбинации признаков. $Y = (L^T U_1, \ldots, L^T U_n)^T$.

Определение. *Совместное распределение матрицы*

$$S = (S_{ij})_{i,j=1}^p = \sum_{r=1}^n U_r U_r^T = (Y_i^T Y_j))_{i,j=1}^p$$

называется центральным распределением Уишарта $W_p(n, \Sigma)$.

Например, при $p = 1$ имеем $S = U_1^2 + \ldots + U_n^2$, которая имеет распределение $\sigma^2 \chi^{\mathbf{2}}(n)$. При $p = 2$ матрица S имеет вид

$$S = \begin{bmatrix} U_{11}^2 + \ldots + U_{1n}^2 & U_{11}U_{21} + \ldots + U_{1n}U_{2n} \\ U_{21}U_{11} + \ldots + U_{2n}U_{1n} & U_{21}^2 + \ldots + U_{2n}^2 \end{bmatrix}$$

Далее мы будем обращаться к линейной функции $L^T U_j \sim \mathcal{N}_1(0, \sigma_L^2)$, где для каждого $j = 1, \ldots, n$ $\sigma_L^2 = L^T \Sigma L$. Вектор

$$Y = \begin{bmatrix} L^T U_1 \\ \vdots \\ L^T U_n \end{bmatrix} = \begin{bmatrix} U_1^T L \\ \vdots \\ U_n^T L \end{bmatrix} = \begin{bmatrix} U_{11} & \ldots & U_{p1} \\ \vdots & \ldots & \vdots \\ U_{1n} & \ldots & U_{pn} \end{bmatrix} \cdot \begin{bmatrix} L_1 \\ \vdots \\ L_p \end{bmatrix} = \mathcal{U} L$$

независимых нормально распределенных случайных величин имеет представление

$$Y = L_1Y_1 + \ldots + L_pY_p = \left[\, Y_1| \;\; \ldots \;\; |Y_p \,\right] \cdot \left[\begin{array}{c} L_1 \\ \vdots \\ L_p \end{array}\right].$$

Для доказательства следующего утверждения требуется теорема Фишера-Конрена и ее следствие о распределении хи-квадрат квадратичной формы с идемпотентой матрицей.

Утверждение 1. $\mathcal{U}^TA\mathcal{U} \sim W_p(n,\Sigma) \iff Y^TAY \sim \sigma_L^2\chi^2(n)$.

Доказательство.

$\Longrightarrow$ Пусть $\mathcal{U}^TA\mathcal{U} \sim W_p(n,\Sigma)$. Обозначим через $S = \mathcal{U}^TA\mathcal{U} \sim W_p(n,\Sigma)$, то есть $S = \sum\limits_{r=1}^{n} U_rU_r^T$. Следовательно,

$$L^TSL = \sum_{r=1}^{n}(L^TU_r)(L^TU_r)^T = \sum_{r=1}^{n}(L^TU_r)^2 \sim \sigma_L^2\chi^2(n)\,,$$

с другой стороны, $L^TSL = L^T\mathcal{U}^TA\mathcal{U}L = (\mathcal{U}L)^TA(\mathcal{U}L) = Y^TAY$.

$\Longleftarrow$ Не умаляя общности будем считать $\sigma_L^2 = 1$. Пусть $Y^TAY \sim \chi^2(n)$. Тогда A – идемпотентная матрица ранга n, которая допускает разложение по ортонормированным векторам

$$A = B_1B_1^T + \ldots + B_nB_n^T,$$
$$\mathcal{U}^TA\mathcal{U} = \mathcal{U}^TB_1B_1^T\mathcal{U} + \ldots + \mathcal{U}^TB_nB_n^T\mathcal{U} = V_1V_1^T + \ldots + V_nV_n^T,$$

где $V_j = \mathcal{U}^TB_j = (Y_1^TB_j, \ldots, Y_p^TB_j)^T$. Так как векторы B_j ортонормальны, то V_j независимые нормальные $\mathcal{N}_p$ случайные величины, и утверждение слудует из определения распределения Уишарта.

$$cov(V_1,V_2) = \mathbf{E}V_1V_2^T = \mathbf{E}\left[\begin{array}{c} B_1^TY_1 \\ \vdots \\ B_1^TY_p \end{array}\right](Y_1^TB_2, \ldots, Y_p^TB_2) =$$

$$= \left[\begin{array}{ccc} B_1^T\mathbf{E}Y_1Y_1^TB_2 & \ldots & B_1^T\mathbf{E}Y_1Y_p^TB_2 \\ \ldots & \vdots & \ldots \\ B_1^T\mathbf{E}Y_pY_1^TB_2 & \ldots & B_1^T\mathbf{E}Y_pY_p^TB_2 \end{array}\right] = \left[\begin{array}{ccc} \sigma_{11}B_1^T\mathbf{I}B_2 & \ldots & \sigma_{1p}B_1^T\mathbf{I}B_2 \\ \ldots & \vdots & \ldots \\ \sigma_{p1}B_1^T\mathbf{I}B_2 & \ldots & \sigma_{pp}B_1^T\mathbf{I}B_2 \end{array}\right] = \mathbf{0}\,.$$

Утверждение 2. Матрицы $\mathcal{U}^T A_1 \mathcal{U}$ и $\mathcal{U}^T A_2 \mathcal{U}$ имеют независимые распределения W_p тогда и только тогда, когда $Y^T A_1 Y$ и $Y^T A_2 Y$ имеют независимые χ^2 распределения для любого L. $\mathcal{U}^T B$ и $\mathcal{U}^T A \mathcal{U}$ распределены независимо как $\mathcal{N}_p$ и W_p в случае, если $Y^T B$ и $Y^T A Y$ распределены независимо как $\mathcal{N}_1$ и χ^2 для любого L.

Утверждение 3 (совместное распределение выборочного среднего и дисперсионной матрицы). Пусть $U_1, \ldots, U_n$ независимые $\mathcal{N}_p(\mu, \Sigma)$ случайные величины, $\bar{U}$ выборочное среднее,

$$(S_{ij}) = \sum U_i U_i^T - n\bar{U}\bar{U}^T$$

матрица сумм квадратов и произведений уклонений. Тогда

$$\bar{U} \ \sim \ \mathcal{N}_p\left(\mu, \frac{1}{n}\Sigma\right), \qquad (S_{ij}) \sim W_p(n-1, \Sigma),$$

а также $\bar{U}$ и (S_{ij}) независимы.

Доказательство. Рассмотрим независимые $\mathcal{N}_1(L^T\mu, L^T\Sigma L)$ случайные величины $L^T U_1, \ldots, L^T U_n$.

$$\frac{1}{n}\sum L^T U_j = L^T\bar{U} \ \sim \ \mathcal{N}_1\left(L^T\mu, \frac{1}{n}L^T\Sigma L\right),$$

$$\sum_{j=1}^{n}(L^T U_j)^2 - n(L^T\bar{U})^2 =$$

$$= L^T\left(\sum_{j=1}^{n} U_j U_j^T - n\bar{U}\bar{U}^T\right)L = L^T(S_{ij})L \ \sim \ (L^T\Sigma L)\chi^2(n-1).$$

Таким образом из независимости $L^T\bar{U}$ и $L^T(S_{ij})L$ для любого L (утв.2) следует независимость $\bar{U}$ и (S_{ij}) и соответствующий вид распределений.

Утверждение 4. Пусть $U_1, \ldots, U_n$ независимые $\mathcal{N}_p(0, \Sigma)$ величины, $S \sim W_p(k, \Sigma)$, B матрица размерности q на p. Тогда $BSB^T \sim W_q(k, B\Sigma B^T)$.

Доказательство.

$$S = \sum_{j=1}^{k} U_j U_j^T, \implies BSB^T = B(\sum_{j=1}^{k} U_j U_j^T)B^T.$$

Так как $BU_j \sim \mathcal{N}_q(0, B\Sigma B^T)$ независимы, то,

$$BSB^T = \sum_{j=1}^{k} (BU_j)(BU_j)^T \ \sim \ W_q(k, B\Sigma B^T),$$

Утверждение 5. Пусть существует обратные матрицы (S^{ij}) и (σ^{ij}), $S \sim W_p(k, \Sigma)$. Тогда статистика

$$\frac{\sigma^{pp}}{S^{pp}} \ \sim \ \chi^2(k-p+1) \tag{52}$$

не зависит от (S_{ij}), $i, j = 1, \ldots, p-1$, и для любого L

$$\frac{L^T\Sigma^{-1}L}{L^T S^{-1}L} \ \sim \ \chi^2(k-p+1). \tag{53}$$

Доказательство. Рассмотрим независимые центрированные $U_1, \ldots, U_k$, образующие матрицу $S = \sum\limits_{i=}^{k} U_i U_i^T \sim W_p(k, \Sigma)$. Обозначим остаток через $\eta_j = U_{pj} - \beta_1 U_{1j} - \ldots - \beta_{p-1} U_{p-1,j}$, где β_i коэффициенты наилучшего линейного предсказания переменной U_{pj} по $p-1$ переменным $U_{1j}, \ldots, U_{p-1,j}$, $E\eta_j = 0$. Помним, что остаток не коррелирован ни с одной вычитаемой переменной. Согласно (12), $D\eta_j = \frac{1}{\sigma^{pp}}$. Используя (33) и заменив в (36) n на k, p на $p-1$ и учитывая центрированность данных, имеем

$$\sum_{j=1}^{k} \frac{\eta_j^2}{1/\sigma^{pp}} \ \sim \ \chi^2(k-(p-1)), \iff$$

$$\iff \sum_{j=1}^{k} \eta_j^2 \overset{(45)}{=} \frac{1}{S^{pp}} \ \sim \ \frac{1}{\sigma^{pp}}\chi^2(k-p+1).$$

Вместо компоненты U_{pj} можно было выбрать любую другую компоненту, поэтому

$$\frac{\sigma^{ii}}{S^{ii}} \ \sim \ \chi^2(k-p+1). \tag{54}$$

Для доказательства второго предложения рассмотрим BSB^T, где B ортогональная матрица, у которой первая строка пропорциональна вектору L. Из утверждения 4

$$BSB^T \sim W_p(k, B\Sigma B^T),\ (BSB^T)^{-1} = BS^{-1}B^T,$$
$$(B\Sigma B^T)^{-1} = B\Sigma^{-1}B^T\,.$$

Первый диагональный элемент $BS^{-1}B^T$ пропорционален $L^TS^{-1}L$, первый диагональный элемент в матрице $B\Sigma^{-1}B^T$ пропорционален $L^T\Sigma^{-1}L$ с одним и тем же коэффициентом пропорциональности. Применяя (54) при $i = 1$, получаем

$$\frac{L^T\Sigma^{-1}L}{L^TS^{-1}L} \sim \chi^2(k-p+1). \tag{55}$$

5.2. Обобщенная статистика Хотеллинга

Определение. $S \sim W_p(k, \Sigma)$, $d \sim \mathcal{N}_p(\mu, c^{-1}\Sigma)$ - независимые случайные величины. Тогда обобщенная статистика Хотеллинга определяется как

$$T^2 = ckd^TS^{-1}d = \frac{kd^TS^{-1}d}{d^T\Sigma^{-1}d}cd^T\Sigma^{-1}d\,. \tag{56}$$

Будем рассматривать наиболее важный частный случай $\mu = 0$.

Предложение 1.

$$\frac{k-p+1}{p}\cdot\frac{T^2}{k} \sim F(p, k-p+1)\,. \tag{57}$$

Согласно (55),

$$\frac{d^T\Sigma^{-1}d}{d^TS^{-1}d} \sim \chi^2(k-p+1)\,.$$

Это распределение не зависит от d, а поскольку $d \sim \mathcal{N}_p(0, c^{-1}\Sigma)$, то

$$cd^T\Sigma^{-1}d \sim \chi^2(p)\,.$$

Распределения независимы, поэтому

$$F = \frac{cd^T\Sigma^{-1}d/p}{\frac{d^T\Sigma^{-1}d}{d^TS^{-1}d}/(k-p+1)} = \frac{k-p+1}{p}\cdot cd^TS^{-1}d \;\sim\; F(p,k-p+1)\,. \quad (58)$$

□

5.3. Критерии, основанные на статистике Хотеллинга

Статистика для расстояния Махаланобиса

Рассмотрим при объеме выборки N выражение для выборочной ковариационной матрицы

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T, \text{ где } \bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i, \quad (59)$$

и, обозначив через $d = x - \bar{x}$, введем расстояние Махаланобиса от вектора x до $x_1,\ldots,x_N$

$$D^2 = (x-\bar{x})^T\hat{\Sigma}^{-1}(x-\bar{x}) = d^T\hat{\Sigma}^{-1}d\,. \quad (60)$$

$\mathbf{D}(x-\bar{x}) = \Sigma + \frac{1}{N}\Sigma = \frac{N+1}{N}\Sigma$, отсюда $c = \frac{N}{N+1}$.

$$S = \sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T \;\sim\; W_p(N-1,\Sigma), \text{ то есть } k = N-1,$$

$$\hat{\Sigma} = \frac{S}{N-1}, \;\Longrightarrow \hat{\Sigma}^{-1} = (N-1)S^{-1},$$

$$d^TS^{-1}d = \frac{D^2}{N-1},$$

Подставляем эти значения в (58) и получаем:

$$\frac{k-p+1}{p}\cdot cd^TS^{-1}d = \frac{N-p}{p}\cdot\frac{N}{N+1}\cdot\frac{D^2}{N-1}\,.$$

Таким образом,

$$\frac{(N-p)N}{(N^2-1)p}D^2 \;\sim\; F(p,N-p)\,. \quad (61)$$

Проверка гипотезы $H_0 : \mu = \mu_0$

На основании статистики Хотеллинга можно проверить гипотезу $H_0 : \mu = \mu_0$ о равенстве вектора средних μ некоторому наперед заданному вектору μ_0. Как и ранее обозначим через $\bar{X}$ и $\hat{\Sigma} = \frac{S}{N-1}$ соответственно вектор выборочных средних и выборочную ковариационную матрицу. Основная статистика имеет вид:

$$T_p^2 = N(\bar{X} - \mu)^T \hat{\Sigma}^{-1} (\bar{X} - \mu),$$

Используем (58), где $d = \bar{X} - \mu \ \sim \ \mathcal{N}(0, \frac{1}{N}\Sigma)$, $c = N$, $S^{-1} = \frac{1}{N-1}\hat{\Sigma}^{-1}$, $k = N - 1$. Тогда

$$F = \frac{k-p+1}{p} \cdot cd^T S^{-1} d =$$
$$= \frac{N-1-p+1}{p} \cdot \frac{N(\bar{X}-\mu)^T \hat{\Sigma}^{-1}(\bar{X}-\mu)}{(N-1)} =$$
$$= \frac{N-p}{p(N-1)} T_p^2 \ \sim \ F(p, N-p)\,.$$

Проверка гипотезы $H_0 : \mu_1 = \mu_2$

Аналогично можно построить критерий для разности между средними значениями. Пусть имеются две выборки объема N_1 и N_2, полученные на p-мерных генеральных совокупностях. Выборочные средние обозначим через $\bar{X}_1$ и $\bar{X}_2$.

$S^{(1)}$ и $S^{(2)}$ – внутривыборочные матрицы сумм и произведений с числом степеней свободы $N_1 - 1$ и $N_2 - 1$;

$S^{(0)} = S^{(1)} + S^{(2)}$ – объединенная матрица с числом степеней свободы $N_1 + N_2 - 2$. $\hat{\Sigma} = \frac{S^{(0)}}{N_1+N_2-2}$, $\hat{\Sigma}^{-1} = (S^{(0)})^{-1}(N_1 + N_2 - 2)$. Статистика Махаланобиса определяется как

$$D_p^2 = (\bar{X}_1 - \bar{X}_2)^T \hat{\Sigma}^{-1} (\bar{X}_1 - \bar{X}_2),$$

Обозначим через $d = \bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(0, c^{-1}\Sigma)$, где

$$c = \frac{N_1 N_2}{N_1 + N_2},\ k = N_1 + N_2 - 2.$$

Используем $S = S^{(0)} \sim W_p(N_1 + N_2 - 2, \Sigma)$, тогда из (58)

$$\begin{aligned} F = \frac{k-p+1}{p} \cdot c d^T S^{-1} d = \\ = \frac{N_1 + N_2 - 2 - p + 1}{p} \cdot \frac{N_1 N_2}{N_1 + N_2} (\bar{X}_1 - \bar{X}_2)^T (S^{(0)})^{-1} (\bar{X}_1 - \bar{X}_2) = \\ = \frac{(N_1 + N_2 - p - 1)}{p} \cdot \frac{N_1 N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} D_p^2 \\ \sim F(p, N_1 + N_2 - p - 1)\,. \end{aligned}$$

5.4. Анализ выбросов при известных и неизвестных параметрах

В разделе 2.6 было показано, что если случайный вектор X с p компонентами имеет многомерное нормальное распределение с вектором средних μ и ковариационной матрицей Σ, то

$$\chi^2 = (X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2(p)\,. \tag{62}$$

При известных параметрах эта статистика может использоваться для проверки аномальности случайного вектора X. Если при наблюдении X, которому соответствует значение статистики χ^2_*, доверительный уровень вероятности

$$p = P\{\chi^2 > \chi^2_*\} < \alpha$$

меньше заранее заданного уровня значимости α, то это наблюдение считается аномальным, а его координаты должны быть проверены на наличие ошибок.

В большинстве случаев параметры не известны и оцениваются (59) по выборке $x_1, \ldots, x_N \sim \mathcal{N}(\mu, \Sigma)$. Выборочным аналогом величины (62)

является расстояние Махаланобиса (60), для проверки значимости которого используется статистика (61).

Для каждого вектора $x_i,\ i = 1, \ldots, n$, по всем $N = n - 1$ векторам наблюдений, исключая x_i, вычисляются вектор средних $\bar{x}_i$, выборочная ковариационная матрица $\hat{\Sigma}_i$, расстояния Махаланобиса D_i^2 между x_i и $\bar{x}_i$ и соответствующие p-значения p_i.

Если все $p_i > \alpha$, то считается, что в выборке нет выбросов. Если некоторые $p_i < \alpha$, то наблюдение с наименьшим p_i исключается, и процедура повторяется для оставшихся $n - 1$ наблюдений.

5.5. Задание по теме „Анализ выбросов“ и „Статистика Хотелинга“

Исследовать на наличие выбросов данные об американских городах (табл.2) и Проверить однородность статистических данных в штатах 2 и 3 (табл.2) по набору признаков: варианты 1 (1, 2, 3), 2 (4, 5, 6), 3 (1, 3, 5), 4 (2, 4, 6), 5 (1, 4, 7), 6 (2, 5, 7), 7 (3, 4, 7).

6. Дисперсионный анализ

6.1. Линейная модель с фиксированными эффектами

Предполагаем, что у нас есть r нормально распределенных популяций $\mathcal{N}(\mu_1, \sigma), \ldots, \mathcal{N}(\mu_r, \sigma)$, где через μ_i обозначены генеральные средние внутри каждой популяции. Переменная x_{ij} означает j-е наблюдение в i-й подпопуляции. Это предположение можно записать в виде:

$$x_{ij} = \mu_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i, \tag{63}$$

где "ошибки" e_{ij} независимы и распределены нормально по $\mathcal{N}(0, \sigma)$. Соотношения (63) представляют собой одну из форм модели дисперсионного анализа.

Во многих случаях желательно выразить i-е среднее μ_i в виде суммы генерального среднего μ и дифференциальных или главных эффектов α_i, определяемых для каждой подпопуляции. Перепишем модель однофакторного дисперсионного анализа в виде

$$x_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i. \tag{64}$$

Наилучшими оценками параметров модели (64) в смысле метода наименьших квадратов являются

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij}, \quad \hat{\alpha}_i = \bar{x}_{i\cdot} - \bar{x} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} - \bar{x}.$$

Ошибкой в этой модели является выражение

$$Q_2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu} - \hat{\alpha}_i)^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2.$$

Ее число степеней свободы равно $\nu_2 = n - r$. Принятие гипотезы H_0 : $\alpha_1 = \ldots = \alpha_r = 0$ означает справедливость эквивалентной гипотезы H_0 :

$\mu_1 = \ldots = \mu_r = \mu$, что все средние по подпопуляциям равны генеральному среднему. Для проверки этой гипотезы используется критерий Фишера. Приведем обоснование этого критерия в рамках общей линейной модели.

При справедливости нулевой гипотезы $H_0 : \alpha_1 = \ldots = \alpha_r = 0$ имеет место усеченная модель вида

$$x_{ij} = \mu + e_{ij},\ i = 1, \ldots, r;\ j = 1, \ldots, n_i. \tag{65}$$

Наилучшей оценкой параметра модели (65) в смысле метода наименьших квадратов является

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij}.$$

Ошибкой $\widetilde{Q_2}$ в этой модели является выражение

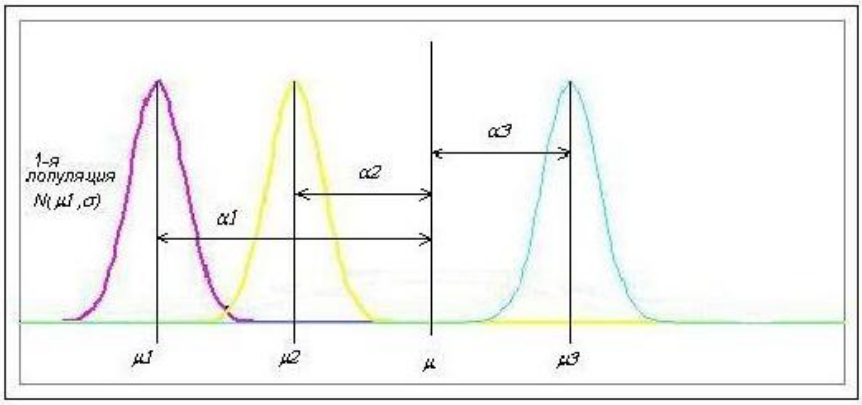


Рис. 1. Модель с фиксированными эффектами. Проверяется гипотеза $H_0 : \mu_1 = \ldots = \mu_r$ или $H_0 : \alpha_1 = \ldots = \alpha_r = 0$; значения $\mu_1, \ldots, \mu$ выбираются по плану.

$$\widetilde{Q_2} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = Q.$$

Ее число степеней свободы равно $\widetilde{\nu_2} = n - 1$. Статистика критерия правдоподобия имеет вид:

$$F = \frac{(\widetilde{Q_2} - Q_2)/(\widetilde{\nu_2} - \nu_2)}{Q_2/\nu_2} = \frac{(Q - Q_2)/(n - 1 - n + r)}{Q_2/(n - r)} = \frac{Q_1/(r - 1)}{Q_2/(n - r)}. \tag{66}$$

Естественно, что ошибка $\widetilde{Q_2}$ в усеченной модели (65) больше, чем ошибка в расширенной модели (64). Если разность $\widetilde{Q_2}-Q_2$ мала, соответственно мало значение статистики F, то усеченная модель не хуже основной модели. Если разность $\widetilde{Q_2}-Q_2$ велика, то соответствие усеченной модели реальным данным гораздо хуже, поэтому влияние фактора, разделяющего наблюдения по подпопуляциям, значимо. Степень значимости определяется при помощи доверительного уровня вероятности

$$\alpha_* = P\{F > F_*\},$$

где F_* наблюдаемое значение статистики F. При $\alpha_* < \alpha = 0.05$ гипотеза $H_0 : \alpha_1 = \ldots = \alpha_r = 0$ отвергается.

Модель (64) иначе называется моделью с фиксированными эффектами. Это означает, что при повторении эксперимента будут рассматриваться выборки из тех же самых подпопуляций.

6.2. Матричная модель дисперсионного анализа

Модель однофакторного дисперсионного анализа (64)

$$x_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i, \quad \text{где } \sum_{i=1}^{r} \alpha_i = 0,$$

можно записать в матричном виде:

$$Y = X\beta + \epsilon,$$

где $Y = (x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{r1}, \ldots, x_{rn_r})'$ –вектор наблюдений, $\sum_{i=1}^{r} n_i = n$, $\beta = (\mu, \alpha_1, \ldots, \alpha_{r-1})'$ – вектор параметров и матрица плана

размерности n на r имеет вид:

$$X = \left[\begin{array}{ccccc} 1 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & \dots & 0 \\ \hline 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & \dots & 0 \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline 1 & -1 & -1 & \dots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & -1 & -1 & \dots & -1 \end{array}\right], \qquad \beta = \left[\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_{r-1} \end{array}\right]$$

Оценки параметров по методу наименьших квадратов имеют вид:

$$\hat{\beta} = (X'X)^{-}X'Y\,.$$

Для проверки гипотезы $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$ используют модель с ограничением на параметры при помощи матрицы H размерности r (число параметров) на $s = r - 1$

$$H'\beta = \theta_0, \quad \text{где} \quad H' = \left[\begin{array}{ccccc} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{array}\right] \quad \text{и} \quad \theta_0 = \left[\begin{array}{c} 0 \\ 0 \\ \dots \\ 0 \end{array}\right]$$

Можно показать, что $Z = H'\hat{\beta} \ \sim \ \mathcal{N}_s(H'\beta, \sigma^2 D)$, где $D = H'(X'X)^{-}H$ и $R_0^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \ \sim \ \sigma^2\chi^2(n-r)$ распределены независимо. Кроме того $(Z - \theta_0)'D^{-1}(Z - \theta_0) = R_1^2 - R_0^2 \ \sim \ \sigma^2\chi^2(s)$, где $R_1^2 = (Y - X\beta^*)'(Y - X\beta^*)$, β^* оценка параметров усеченной модели. Тогда при справедливости нулевой гипотезы

$$F = \frac{n-r}{s}\frac{R_1^2 - R_0^2}{R_0^2} \ \sim \ F(s, n-r)\,.$$

6.3. Дифференцирование по вектору параметров

$$A\beta = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_m \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{m} a_{1i}\beta_i \\ \dots \\ \sum\limits_{i=1}^{m} a_{ni}\beta_i \end{bmatrix}$$

$$\frac{\partial A\beta}{\partial \beta} = \begin{bmatrix} \frac{\partial A\beta}{\partial \beta_1} & \dots & \frac{\partial A\beta}{\partial \beta_m} \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} = A$$

Для получения оценок МНК дифференцируем квадратичную форму.

$$\mathcal{L}_1 = (A\beta)'(A\beta) = \left(\sum_{i=1}^{m} a_{1i}\beta_i\right)^2 + \ldots + \left(\sum_{i=1}^{m} a_{ni}\beta_i\right)^2 .$$

Система нормальных уравнений имеет вид

$$\begin{cases} \frac{\partial \mathcal{L}_1}{\partial \beta_1} = 2\left(a_{11}\left(\sum\limits_{i=1}^{m} a_{1i}\beta_i\right) + \ldots + a_{n1}\left(\sum\limits_{i=1}^{m} a_{ni}\beta_i\right)\right) = 0, \\ \dots \\ \frac{\partial \mathcal{L}_1}{\partial \beta_m} = 2\left(a_{1m}\left(\sum\limits_{i=1}^{m} a_{1i}\beta_i\right) + \ldots + a_{nm}\left(\sum\limits_{i=1}^{m} a_{ni}\beta_i\right)\right) = 0, \end{cases}$$

$$\frac{\partial (A\beta)'(A\beta)}{\partial \beta} = 0 \Longleftrightarrow \quad 2\frac{\partial (A'\beta)}{\partial \beta} A\beta = 2A'A\beta = 0\,.$$

При ОМК с ограничением на параметры дифференцируется выражение

$$\mathcal{L}_2 = \lambda'(1,s)H'(s,m)\beta(m,1) =$$

$$= (\lambda_1, \dots, \lambda_s) \begin{bmatrix} h_{11} & \dots & h_{m1} \\ \dots & \dots & \dots \\ h_{1s} & \dots & h_{ms} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_m \end{bmatrix} =$$

$$= (\lambda_1, \dots, \lambda_s) \begin{bmatrix} \sum\limits_{i=1}^{m} h_{i1}\beta_i \\ \dots \\ \sum\limits_{i=1}^{m} h_{is}\beta_i \end{bmatrix} = \sum_{j=1}^{s} \lambda_j \left(\sum_{i=1}^{m} h_{ij}\beta_i \right) .$$

Система нормальных уравнений

$$\begin{cases} \frac{\partial \mathcal{L}_2}{\partial \beta_1} = \sum\limits_{j=1}^{s} \lambda_j h_{1j} = 0, \\ \dots \\ \frac{\partial \mathcal{L}_2}{\partial \beta_m} = \sum\limits_{j=1}^{s} \lambda_j h_{mj} = 0, \end{cases} \iff$$

$$\iff \begin{bmatrix} h_{11} & \dots & h_{1s} \\ \dots & \dots & \dots \\ h_{m1} & \dots & h_{ms} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \dots \\ \lambda_s \end{bmatrix} = H\lambda = \mathbf{0} .$$

6.4. Вспомогательные задачи

Определение. A^- - обобщенная обратная матрица, такая, что для любого вектора Y, при котором система $AX = Y$ совместна, $X = A^-Y$ является ее решением.

A^- существует $\iff AA^-A = A \iff A^-AA^- = A^-$.

Обобщенную обратную матрицу, иногда ее называют псевдообратной, можно вычислить через сингулярное разложение матрицы $A = U\Sigma V^T$, где Σ диагональная матрица с сингулярными числами на главной диагонали, матрицы U, V состоят из соответственно левых и правых собственных векторов (с.в. матриц AA^T, A^TA). Сингулярные числа вычисляются как квадратные корни из собственных чисел матриц AA^T, A^TA. Для псевдообратной матрицы справедливо $A^- = V\Sigma^{-1}U^T$, где у матрицы Σ^{-1} на диагонали обратные к ненулевым сингулярным числам.

Задача 1. *Пусть A — симметричная матрица n на n ранга r, B — матрица m на n, BY — m линейных функций от $Y = (y_1, \ldots, y_n)^T$, y_i независимые нормально распределенные случайные величины. Для независимости BY и Y^TAY достаточно условия $BA = \mathbf{0}_{m,n}$, где $\mathbf{0}_{m,n}$ матрица m на n из нулей.*

Используем спектральное разложение симметричной матрицы A ранга r:

$$A = \lambda_1 P_1 P_1^T + \ldots + \lambda_r P_r P_r^T, \quad \lambda_i \neq 0,\ i = 1, \ldots, r. \tag{67}$$

Пусть $BA = \mathbf{0}$. Умножим (67) слева на B и справа для примера на P_1.

$$BAP_1 = \lambda_1 BP_1 P_1^T P_1 + \ldots + \lambda_r BP_r P_r^T P_1 = \mathbf{0}_{m,1}.$$

Из ортогональности собственных векторов получаем $BP_1 = \mathbf{0}_{m,1}$. Аналогично для остальных $BP_i = \mathbf{0}_{m,1}$, т.е. линейные функции BY не зависят от функций P_i^TY (если $a^Tb = 0$, то для Y с независимыми компонентами $\mathbb{E}(a^TY)(Y^Tb) = a^T\mathbf{I}b = a^Tb = 0$) и не зависят от Y^TAY, так как

$$Y^TAY = \sum \lambda_i (P_i^TY)^2.$$

Задача 2.

а) $X = X(X'X)^-X'X$,

б) $I - X(X'X)^-X'$ идемпотентна.

$$\text{а)}\quad G = X(I - (X'X)^-X'X),$$

$$\Longrightarrow \quad G'G = (I - (X'X)^-X'X)'X'X(I - (X'X)^-X'X) =$$

$$= (I - (X'X)^-X'X)'(X'X - X'X(X'X)^-X'X) =$$

$$= (I - (X'X)^-X'X)'(X'X - X'X) = 0, \quad \Longrightarrow \quad G = 0.$$

$$\text{б)}\quad (I - X(X'X)^-X')(I - X(X'X)^-X') =$$

$$= I - X(X'X)^-X' - X(X'X)^-X' + \underbrace{X(X'X)^-X'X}_{X}(X'X)^-X' =$$

$$= I - X(X'X)^-X'.$$

Задача 3. *Пусть* $\beta = (\beta_1, \ldots, \beta_m)'$ *вектор параметров,* X *- матрица плана ранга* r*,* $(X'X)^-$ *обобщенная обратная,* $\hat{\beta} = (X'X)^- X'Y$*,* $\hat{\beta}$ *решение уравнения* $(X'X)\beta = X'Y$*. Матрица* H' *размерности* s *на* m *ранга* s*. Показать, что распределены независимо* $Z = H'\hat{\beta}$ *и*

$$R_0^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta}).$$

Введем матрицу C такую, что $H = X'C$. Тогда

$$H'\hat{\beta} = \overbrace{C'X(X'X)^-X'}^{B}Y = BY.$$

$$R_0^2 = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'\underbrace{X'X\hat{\beta}}_{X'Y} = Y'Y - Y'X\hat{\beta}, \qquad (68)$$

$$R_0^2 = Y'Y - Y'X(X'X)^-X'Y = Y'(\underbrace{I - X(X'X)^-X'}_{A})Y = Y'AY\,.$$

Таким образом $BA = C'X(X'X)^-X'(I - X(X'X)^-X') =$

$$= C'(X(X'X)^-X' - \underbrace{X(X'X)^-X'X}_{X}(X'X)^-X') = \mathbf{0}\,.$$

Задача 4.

а) $Z = H'\hat{\beta} \sim \mathcal{N}_s(H'\beta, \sigma^2 D)$*, где* $D = H'(X'X)^-H$ *и*

б) $R_0^2 \sim \sigma^2\chi^2(n-r)$*, где* r *– ранг матрицы плана* X*.*

а) Действительно, $E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' =$

$$= E((X'X)^-X'(Y - EY))((X'X)^-X'(Y - EY))' =$$

$$= (X'X)^-X'\sigma^2 I((X'X)^-X')' =$$

$= \sigma^2(X'X)^-X'X(X'X)^- = \sigma^2(X'X)^-$. Отсюда $D = H'(X'X)^-H$.

б) $R_0^2 = Y'(I - X(X'X)^-X')Y$, матрица $I - X(X'X)^-X'$ идемпотентна,

$\overset{\text{раздел 5.1}}{\Longrightarrow} R_0^2 \sim \sigma^2\chi^2(n-r)$, где $n - r = TrI - TrX(X'X)^-X' =$

$= n - Tr(X'X)^-X'X = n - rang(X'X) = n - rang(X) = n - r$.

Задача 5. При дифференцировании $(Y - X\beta)'(Y - X\beta)$ по β при условии $H'\beta = \theta_0$ получена система уравнений (разд. 6.3):

$$\begin{cases} X'X\beta + H\lambda = X'Y, \\ H'\beta = \theta_0 \end{cases} \tag{69}$$

λ^* и β^* – ее решение. Показать, что

$$R_1^2 = (Y - X\beta^*)'(Y - X\beta^*) = (Z - \theta_0)'D^{-1}(Z - \theta_0) + R_0^2 .$$

Заметим, что

$$\begin{aligned} R_1^2 &= (Y - X\beta^* + X\hat{\beta} - X\hat{\beta})'(Y - X\beta^* + X\hat{\beta} - X\hat{\beta}) = \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta} - \beta^*)'X'X(\hat{\beta} - \beta^*)', \\ &\text{так как } (\hat{\beta} - \beta^*)'X'(Y - X\hat{\beta}) = 0 \text{ по определению } \hat{\beta} . \end{aligned}$$

Достаточно показать, что

$$(\hat{\beta} - \beta^*)'X'X(\hat{\beta} - \beta^*)' = (Z - \theta_0)'D^{-1}(Z - \theta_0).$$

Пусть $H = X'XC$, тогда $D = H'(X'X)^-H = C'X'X(X'X)^-X'XC = C'X'XC$. Вычитаем из $X'X\hat{\beta} = X'Y$ уравнение (69) $X'X\beta^* + H\lambda^* = X'Y$.

$$X'X(\hat{\beta} - \beta^*) = H\lambda^* = X'XC\lambda^*, \quad \hat{\beta} - \beta^* = (X'X)^-X'XC\lambda^*,$$

$$(\hat{\beta} - \beta^*)'X'X(\hat{\beta} - \beta^*) = \underbrace{\lambda^{*'}C'X'X(X'X)^-}_{(\hat{\beta}-\beta^*)'} X'XC\lambda^* =$$

$$= \lambda^{*'}C'X'XC\lambda^* = \lambda^{*'}D\lambda^*$$

$$Z - \theta_0 = H'\hat{\beta} - H'\beta^* = H'(\hat{\beta} - \beta^*) = C'X'X(X'X)^-X'XC\lambda^* =$$

$$= C'X'XC\lambda^* = D\lambda^*$$

$$\Longrightarrow \lambda^* = D^{-1}(Z - \theta_0),$$

$$\Longrightarrow \lambda^{*'}D\lambda^* = (Z - \theta_0)'D^{-1}DD^{-1}(Z - \theta_0) = (Z - \theta_0)'D^{-1}(Z - \theta_0) .$$

6.5. Поправки Бонферони

В методе LSD (the least significant difference — наименее значимое различие) применяются стандартные t-критерии ко всем возможным парам средних групп. Пусть имеется r выборок $\mathcal{N}(\mu_i, \sigma)$ x_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$, $n = \sum\limits_{i=1}^{r} n_i$, ранее было показано, что

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n-r), \quad \text{где } s^2 = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - x_{..})^2$$

Поправки Бонферрони применяются для того, чтобы частота ложноположительных результатов с поправкой на эффект множественных сравнений не превышала заданное значение.

Пусть имеются нулевые гипотезы $\mathcal{H}_1, \ldots, \mathcal{H}_m$ с соответствующими p-значениями: $p_1, \ldots, p_m$. Например, $m = C_r^2$, если проверяются гипотезы о равенстве всех попарных средних. Определим как вероятность совершить хотя бы одну ошибку в выводе из m выводов. Если осуществляется n независимых испытаний с вероятностью ошибки α в каждом, то $\tilde{\alpha} = 1-(1-\alpha)^n$. В случае зависимых испытаний $\tilde{\alpha} \leqslant 1 - (1-\alpha)^n$. Заметим, что по неравенству Буля

$$\tilde{\alpha} = 1 - (1 - n\alpha + C_n^2\alpha^2 - \ldots) \leqslant n \cdot \alpha.$$

Таким образом, если мы хотим, чтобы вероятность наличия хотя бы одного неверного вывода из m была равна 0.05, то достаточно установить вероятность неверного отклонения нулевой гипотезы равной $0.05/m$ для каждого вывода. Такой метод называется поправкой Бонферрони (англ.). Если гипотезы имеют различные доверительные уровни вероятностей, то достаточно отвергнуть гипотезы, имеющие $\alpha < \tilde{\alpha}/m$.

6.6. Метод множественных сравнений

Определение 1. Сравнением параметров $\beta_1, \ldots, \beta_p$ называется линейная функция $\sum\limits_{i=1}^{p} c_i\beta_i$, где $\sum\limits_{i=1}^{p} c_i = 0$.

Например, имеются три градации категориального признака, из которых две подвергнуты некоторому условию, а одна нет. Для проверки этого утверждения можно рассмотреть разность

$$\psi = \frac{\beta_1 + \beta_2}{2} - \beta_3 \,.$$

Обозначим через $\hat{\psi} = \sum\limits_{i=1}^{p} c_i\hat{\beta}_i$ несмещенную оценку сравнения. В матричном виде

$$\psi_{q,1} = C_{q,p}\beta_{p,1}, \quad rank(C) = q,$$
$$\hat{\psi} = C\hat{\beta} = C(X^TX)^{-}X^TY = AY.$$

Ковариационная матрица оценок находится по формуле

$$\Gamma_\psi = \sigma^2 AA^T, B = AA^T,$$

несмещенной оценкой σ^2 является средний квадрат ошибок

$$s^2 = \frac{R_0^2}{n-r}$$

Теорема 1. *(Метод множественных сравнений Шеффе) Если* $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, $rank(X) = r$, *то случайная величина* $\hat{\psi} \sim \mathcal{N}(\psi, \Gamma_\psi)$ *и не зависит от* $R_0^2/\sigma^2 \sim \chi^{\mathbf{2}}(n-r)$. *Поэтому*

$$\frac{(\hat{\psi}-\psi)^T B^{-1}(\hat{\psi}-\psi)}{qs^2} \sim F(q, n-r).$$

Пусть элементы выборки $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma)$ упорядочены в виде $x^{(1)} < \ldots < x^{(n)}$, разность $R = x^{(n)} - x^{(1)}$ называется размахом выборки.

Пусть s^2 является независимой среднеквадратичной оценкой σ^2 с ν степенями свободы. Таким образом $\nu s^2/\sigma^2 = \chi^2_\nu$ не зависит от R. Случайную величину $R/s = q_{n,\nu}$ называют стьюдентизированным размахом.

Метод Тьюки можно применять для получения совместных доверительных утверждений о сравнениях множества параметров $\theta_1, \ldots, \theta_k$ в терминах несмещенных оценок. Ограничением метода Тьюки является требование одинаковых дисперсий для оценок θ_i. Поэтому если нужна классификация по одному признаку, то объемы должны быть равными.

Теорема 2. *(Метод множественных сравнений Тьюки) Пусть* $\hat{\theta}_1, \ldots, \hat{\theta}_k$ *независимы,* $\mathcal{N}(\theta_i, a^2\sigma^2)$, $\frac{\nu s^2}{\sigma^2} = \chi^2_\nu$ *не зависит от* $\{\theta_i\}$, $\hat{\psi} = \sum_{i=1}^{k} c_i\hat{\theta}_i$, $T = aq_{\alpha,k,\nu}$, *где* $q_{\alpha,k,\nu}$ *верхний* α *предел стьюдентизированного размаха. Тогда вероятность того, что все* $\frac{k(k-1)}{2}$ *разностей* $\{\theta_i - \theta_j\}$ *одновременно удовлетворяют неравенствам*

$$\hat{\theta}_i - \hat{\theta}_j - Ts \le \theta_i - \theta_j \le \hat{\theta}_i - \hat{\theta}_j + Ts$$

равна $1 - \alpha$.

Кроме того, существует обобщение для $\hat{\theta}_1, \ldots, \hat{\theta}_k$, не обязательно независимых, но имеющих одинаковые ковариации и дисперсии.

В критерии Пиллая используется верхний α предел стьюдентизированного максимума модулей $M = max|x_i|/s$. Увеличенным размахом R' называется $max(R, M)$.

В тех случаях, когда главный интерес представляют все разности, причем никакой из них не отдается предпочтения, метод Тьюки дает более узкие интервалы, но применим только в случае одинаковых дисперсий. Преимущество метода Шеффе в том, что он не чувствителен к нарушению о предположении нормальности и равенства дисперсий.

6.7. Многомерный дисперсионный анализ

Пусть вектор $Y_i = (y_{i1}, \ldots, y_{in})'$ соответствует n независимым измерениям i-й переменной, $i = 1, \ldots, p$. Для каждого Y_i можно рассмотреть одномерную линейную модель

$$Y_i = X\beta_i + e_i, \quad cov(Y_i) = \sigma_{ii} I, \tag{70}$$

где X – матрица плана ранга $r \leq m < n$, $\beta_i = (\beta_{i1}, \ldots, \beta_{im})'$ – вектор параметров, специфичных для каждой переменной. Через σ_{ij} обозначена ковариация между переменными Y_i и Y_j, $p \leq n - r$, $r < n$. Взятые вместе p линейных моделей (70) составляют *многомерную обобщенную линейную модель:* $Y = X\beta + e$, где

$$Y = \begin{bmatrix} y_{11} & y_{21} & \ldots & y_{p1} \\ y_{12} & y_{22} & \ldots & y_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ y_{1n} & y_{2n} & \ldots & y_{pn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{11} & \beta_{21} & \ldots & \beta_{p1} \\ \beta_{12} & \beta_{22} & \ldots & \beta_{p2} \\ \ldots & \ldots & \ldots & \ldots \\ \beta_{1m} & \beta_{2m} & \ldots & \beta_{pm} \end{bmatrix} \tag{71}$$

X – матрица плана размерности n на m ранга r, e – матрица n на p, строки которой составляют случайную выборку из $\mathcal{N}(0, \Sigma)$.

Оценкой параметров $\beta_i,\ i = 1, \ldots, p$, служит МНК-оценка по одной переменной Y_i вида

$$\hat{\beta}_i = (X'X)^{-}(X'Y_i), \quad i = 1, \ldots, p\,.$$

Для вычисления несмещенных оценок для σ_{ii} и σ_{ij} используются остаточная сумма квадратов (68)

$$R_0^2(i,i) = Y_i'Y_i - Y_i'X\beta_i, \quad \hat{\sigma}_{ii} = \frac{R_0^2(i,i)}{n-r}$$

и остаточная сумма произведений

$$R_0^2(i,j) = Y_i'Y_j - Y_i'X\beta_j, \quad \hat{\sigma}_{ij} = \frac{R_0^2(i,j)}{n-r}\,.$$

$$\mathbf{R_0} = \begin{bmatrix} R_0^2(1,1) & \dots & R_0^2(1,p) \\ \dots & \dots & \dots \\ R_0^2(p,1) & \dots & R_0^2(p,p) \end{bmatrix}$$

называется матрицей остаточных сумм и произведений. Требуется одновременно проверить гипотезы о том, что

$$H'\beta_i = \theta_i,\ i = 1, \dots, p\,, \tag{72}$$

где матрица H' размерности s на m имеет ранг $s \leq r$. Если эта гипотеза выполняется, то получается модель с ограничениями, на основе которой можно найти оценки β_i^* и матрицу остаточных сумм и произведений $\mathbf{R_1}$. Матрица $\mathbf{R_1} - \mathbf{R_0}$ называется матрицей квадратов и произведений, обусловленных отклонением от гипотезы. Таким образом, отклонение от гипотезы может быть определено сравнением матриц $\mathbf{R_1} - \mathbf{R_0}$ и $\mathbf{R_0}$.

Сведем задачу (72) к одномерному случаю, рассмотрев

$$\beta = L_1\beta_1 + \ldots + L_p\beta_p,\ \ \xi = L_1\xi_1 + \ldots + L_p\xi_p,$$

гипотезу $H'\beta = \xi$, множество случайных величин $Y = L_1Y_1 + \ldots + L_pY_p$. Подходящий критерий основывается на двух минимальных суммах квадратов:

$$R_0^2 = min(Y - X\beta)'(Y - X\beta), \quad R_1^2 = \min_{H'\beta=\xi}(Y - X\beta)'(Y - X\beta).$$

При нулевой гипотезе $H'\beta = \xi$ статистики $R_1^2 - R_0^2$ и R_0^2 распределены независимо как центральные хи-квадрат распределения $\sigma_L^2\chi^2(s)$ и $\sigma_L^2\chi^2(n-r)$. Можно непосредственно убедиться в том, что

$$L'\mathbf{R_0}L = R_0^2,\ \ L'\mathbf{R_1}L = R_1^2,\ \ L'(\mathbf{R_1} - \mathbf{R_0})L = R_1^2 - R_0^2,$$

например,

$$R_0^2 = (Y - X\beta)'(Y - X\beta) = \Big[\sum L_i(Y_i - X\beta_i)\Big]' \Big[\sum L_i(Y_i - X\beta_i)\Big] =$$
$$= \sum L_i L_j (Y_i - X\beta_i)'(Y_i - X\beta_i) = \sum L_i L_j R_0^2(i,j) = L'\mathbf{R_0}L\,.$$

Следовательно, матрицы $\mathbf{R_1} - \mathbf{R_0}$ и $\mathbf{R_0}$ имеют независимые распределения Уишарта

$$\mathbf{R_0} \;\sim\; W_p(n-r), \quad \mathbf{R_1} - \mathbf{R_0} \;\sim\; W_p(s).$$

При справедливости нулевой гипотезы (72) для каждого L верна гипотеза $H'\beta = \xi$ и подходящими являются статистики

$$F = \frac{n-r}{s} \cdot \frac{R_1^2 - R_0^2}{R_0^2} \quad \text{или} \quad B = \frac{R_0^2}{R_1^2} = \frac{1}{1 + \frac{sF}{n-r}} \;=\; \frac{L^T\ \mathbf{R_0}\ L}{L^T\ \mathbf{R_1}\ L},$$

последняя имеет бета-распределение. Чем больше F и значимость, тем меньше B. Выберем L так, чтобы статистика B имела наименьшее значение (F наибольшее), тем самым пытаясь опровергнуть гипотезу. Тогда критерий сводится к виду:

$$\lambda = \min_L B = \min_L \frac{L^T\ \mathbf{R_0}\ L}{L^T\ \mathbf{R_1}\ L},$$

или λ является наименьшим корнем характеристического уравнения

$$|\mathbf{R_0} - \lambda\mathbf{R_1}| = 0\,, \quad \text{так как} \quad \mathbf{R_0}L = \lambda\mathbf{R_1}L \iff L^T\mathbf{R_0}L =$$
$$= \lambda L^T\mathbf{R_1}L \iff \lambda = \frac{L^T\ \mathbf{R_0}\ L}{L^T\ \mathbf{R_1}\ L}\,.$$

Один из возможных критериев на собственных числах - Λ-критерий Уилкса:

$$\Lambda = \lambda_1 \cdot \ldots \cdot \lambda_p = \frac{|\mathbf{R_0}|}{|\mathbf{R_1}|}\,.$$

6.8. Двуфакторный дисперсионный анализ с повторениями

Исследуется уровень адреналина у мышей, содержащихся в разных условиях до периода полового созревания, и выкормленных по-разному. Таким образом, имеются два фактора: A - условия вскармливания (без матери и с матерью), B - условия содержания (в изолированных, смежных и общих клетках). Пусть x_{ijk} обозначает значение переменной X (уровень адреналина), полученное при k-м повторении эксперимента в ячейке ij, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$.

Двуфакторная модель с фиксированными эффектами

Модель с фиксированными эффектами имеет вид:

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \tag{73}$$

$i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, μ - генеральное среднее, α_i - дифференциальный эффект фактора A, β_j - дифференциальный эффект фактора B. Величина $(\alpha\beta)_{ij}$ называется взаимодействием факторов. Эта величина учитывает дифференциальный эффект комбинаций i-го уровня фактора A и j-го уровня фактора B, если он не выражается суммой $\alpha_i + \beta_j + \mu$. Ошибки e_{ijk} предполагаются независимыми и нормально распределенными $\mathcal{N}(0, \sigma)$. При ограничениях на параметры

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0,$$

$$\sum_{i=1}^{I} (\alpha\beta)_{ij} = 0, \quad j = 1, 2, \ldots, J; \quad \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0, \quad i = 1, 2, \ldots, I,$$

оценками параметров модели (73) являются:

$\hat{\mu} = \bar{x}$ – общее среднее,

$\hat{\alpha}_i = \bar{x}_{i\cdot} - \bar{x}$, где $\bar{x}_{i\cdot}$ – среднее по i-му значению фактора A;

$\hat{\beta}_j = \bar{x}_{\cdot j} - \bar{x}$, где $\bar{x}_{\cdot j}$ – среднее по j-му значению фактора B;

$\widehat{(\alpha\beta)_{ij}} = \bar{x}_{ij\cdot} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}$, где $\bar{x}_{ij\cdot}$ – среднее в ячейке ij.

Уровень адреналина при разном вскармливании и содержании

вскармл. с матерью 3.35 ± 0.15			вскармл. без матери 4.48 ± 0.15		
отд.клет.ξ_1	сообщ.клет.ξ_2	общ.клет.ξ_3	отд.клет.ξ_1	сообщ.клет.ξ_2	общ.клет.ξ_3
1.9	4	3.2	3.3	6.3	4.6
2.3	4.6	2.6	4	7.2	4.8
2.2	5.7	2.2	5	4.6	4.6
2	5.7	2.6	3.2	7.2	4.4
2.7	4.8	3.2	2.4	3.8	4.5
2.8	4.8	2.5	3.6	4.4	4.2
2.4	5.4	3	3	4.8	4.4
2.7	3.8	3.3	3	5.8	4.3
$\bar{x}_{11} = 2.38$	$\bar{x}_{12} = 4.85$	$\bar{x}_{13} = 2.83$	$\bar{x}_{21} = 3.44$	$\bar{x}_{22} = 5.51$	$\bar{x}_{23} = 4.48$
$\bar{x}_{\cdot 1} = 2.91$	$\bar{x}_{\cdot 2} = 5.18$	$\bar{x}_{\cdot 3} = 3.65$			

Имеем вектор Y размерности IJK, матрицу плана размерности IJK на $r = 1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$. Например, при $I = 3$ и $J = 4$ она имеет вид:

Y	μ	α_1	α_2	β_1	β_2	β_3	$(\alpha\beta)_{11}$	$(\alpha\beta)_{12}$	$(\alpha\beta)_{13}$	$(\alpha\beta)_{21}$	$(\alpha\beta)_{22}$	$(\alpha\beta)_{23}$
y_{11*}	1	1	0	1	0	0	1	0	0	0	0	0
y_{12*}	1	1	0	0	1	0	0	1	0	0	0	0
y_{13*}	1	1	0	0	0	1	0	0	1	0	0	0
y_{14*}	1	1	0	−1	−1	−1	−1	−1	−1	0	0	0
y_{21*}	1	0	1	1	0	0	0	0	0	1	0	0
y_{22*}	1	0	1	0	1	0	0	0	0	0	1	0
y_{23*}	1	0	1	0	0	1	0	0	0	0	0	1
y_{24*}	1	0	1	−1	−1	−1	0	0	0	−1	−1	−1
y_{31*}	1	−1	−1	1	0	0	−1	0	0	−1	0	0
y_{32*}	1	−1	−1	0	1	0	0	−1	0	0	−1	0
y_{33*}	1	−1	−1	0	0	1	0	0	−1	0	0	−1
y_{34*}	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

Вектор параметров: $\beta = (\mu, \alpha_1, \alpha_{I-1}, \beta_1, \ldots, \beta_{J-1}, (\alpha\beta)_{11}, \ldots, (\alpha\beta)_{I-1,J-1})'$.

Общее число наблюдений равно IJK. Остаточная сумма квадратов основ-

ной модели равна

$$Q_R = R_0^2 = \sum_{i,j,k} (x_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - (\alpha\hat{\beta})_{ij})^2$$

имеет число степеней свободы $IJK - IJ$. Для проверки гипотезы $H_0 : \alpha_1 = \alpha_1 = \ldots = \alpha_I = 0$ используем $H'\beta = 0$, матрицу H' размерности $I-1$ на $r = IJ$ ранга $I-1$, вида $H' = \left[\mathbf{0}_{I-1,1}|\mathbf{I}_{I-1,I-1}|\mathbf{0}_{I-1,IJ-I)}\right]$. Остаточная сумма квадратов усеченной модели равна

$$R_1^2 = \sum_{i,j,k} (x_{ijk} - \hat{\mu} - \hat{\beta}_j - (\alpha\hat{\beta})_{ij})^2 = R_0^2 + Q_A,$$

$$\text{где } Q_A = \sum_{i,j,k} (\hat{\alpha}_i)^2 = JK \sum_{i=1}^{I} (\bar{x}_{i\cdot} - \bar{x})^2$$

Отсюда получаем статистику критерия

$$F = \frac{IJK - IJ}{I-1} \cdot \frac{R_1^2 - R_0^2}{R_0^2} = \frac{IJK - IJ}{I-1} \cdot \frac{Q_A}{Q_R} \ \sim \ F(I-1, IJK - IJ)\,.$$

Аналогично строятся остальные статистики.

Таблица двуфакторного дисперсионного анализа

источник дисперсии	сумма квадратов	степени свободы	средний квадрат
фактор A	$Q_A = JK \sum\limits_{i=1}^{I} (\bar{x}_{i\cdot} - \bar{x})^2$	$\nu_A = I - 1$	$MQ_A = \frac{Q_A}{\nu_A}$
фактор B	$Q_B = IK \sum\limits_{j=1}^{J} (\bar{x}_{\cdot j} - \bar{x})^2$	$\nu_B = J - 1$	$MQ_B = \frac{Q_B}{\nu_B}$
взаимо- действие AB	$Q_{AB} = K \sum\limits_{i=1}^{I} \sum\limits_{j=1}^{J} (\bar{x}_{ij\cdot} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$	$\nu_{AB} = (I-1)(J-1)$	$MQ_{AB} = \frac{Q_{AB}}{\nu_{AB}}$
остаток (ошибка)	$Q_R = \sum\limits_{i=1}^{I} \sum\limits_{j=1}^{J} \sum\limits_{k=1}^{K} (x_{ijk} - \bar{x}_{ij\cdot})^2$	$\nu_R = IJ(K-1)$	$MQ_R = \frac{Q_R}{\nu_R}$
полная	$Q_T = \sum\limits_{i=1}^{I} \sum\limits_{j=1}^{J} s \sum\limits_{k=1}^{K} (x_{ijk} - \bar{x})^2$	$\nu_T = IJK - 1$	

Для проверки гипотезы об отсутствии эффекта взаимодействия

$$H_0 : \text{все } (\alpha\beta)_{ij} = 0 \text{ используется статистика } F = \frac{MQ_{AB}}{MQ_R} \sim F(\nu_{AB}, \nu_R)\,.$$

Для проверки гипотезы об отсутствии эффекта A-фактора

$$H_0 : \text{все } \alpha_i = 0 \text{ используется отношение } F = \frac{MQ_A}{MQ_R} \sim F(\nu_A, \nu_R)\,.$$

Для проверки гипотезы об отсутствии эффекта B-фактора

$$H_0 : \text{все } \beta_j = 0 \text{ берется отношение } F = \frac{MQ_B}{MQ_R} \sim F(\nu_B, \nu_R)\,.$$

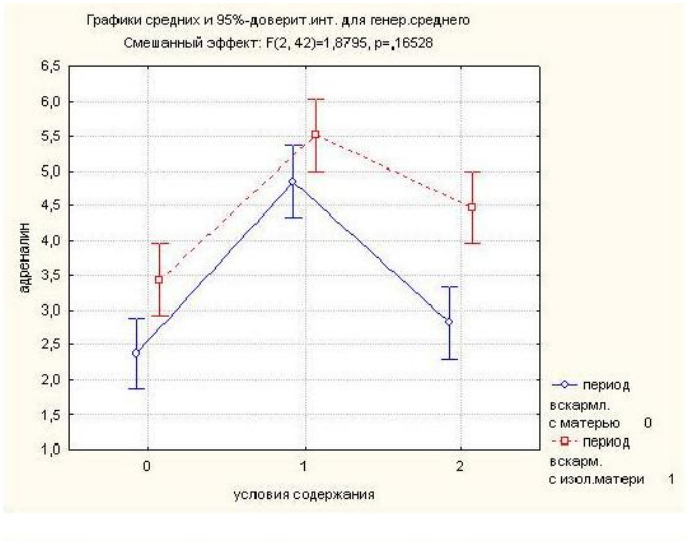


Рис. 2. Иллюстрация отсутствия эффекта взаимодействия факторов вскармливания и содержания.

ПРИМЕР (продолжение) Проверяем гипотезы.

$H_0 : \text{все } \alpha_i = 0$. $p = 0.000003$ – следовательно, влияние фактора вскармливания значимо для уровня адреналина, средний уровень адреналина 3.35 ± 0.15 в группе, до конца вскармливания находящейся с матерью, значимо ниже среднего уровня адреналина 4.48 ± 0.15 в группе, изолированной от матери.

$H_0 : \text{все } \beta_j = 0$. $p < 0.000001$ – следовательно, влияние фактора условия содержания значимо для уровня адреналина: средние 2.91 ± 0.18, 5.18 ± 0.18 и 3.65 ± 0.18 в группах, отличающихся условиями содержания, значимо различаются.

$H_0 : \text{все } (\alpha\beta)_{ij} = 0$. $p = 0.16$ – следовательно, фактор взаимодействия не значим.

6.9. Модель со случайными эффектами

Если подпопуляции выбираются случайно из большого (бесконечного) числа подпопуляций, то фактор относится ко второй модели, которая имеет вид:

$$x_{ij} = \mu + a_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i, \tag{74}$$

где a_i распределены по $\mathcal{N}(0, \sigma_a)$, e_{ij} распределены по $\mathcal{N}(0, \sigma)$, все a_i и e_{ij} в совокупности независимы.

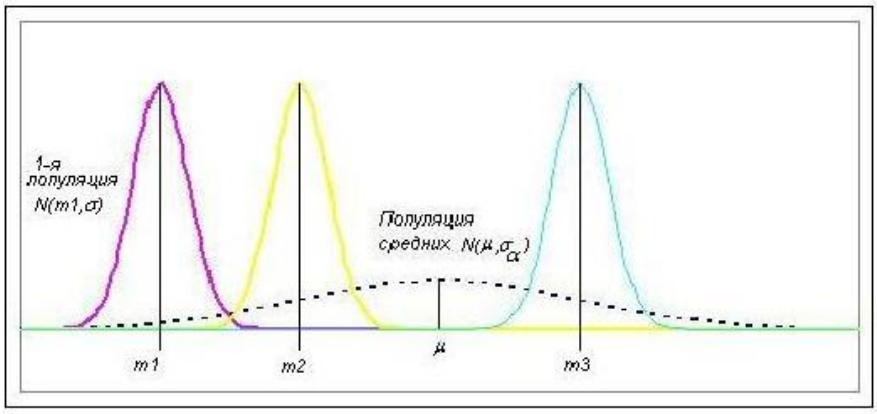


Рис. 3. Модель со случайными эффектами. Проверяется гипотеза $H_0 : \sigma_a^2 = 0, m_1, \ldots, m_r$ выбираются случайно.

При повторении эксперимента мы скорее всего будем иметь дело со случайными выборками из других популяций. В этой модели нас интересует оценка дисперсии σ_a^2 распределения дифференциальных эффектов. Гипотеза $H_0 : \sigma_a^2 = 0$ означает, что фактор не вносит значимого вклада в дисперсию.

Для проверки гипотезы $H_0 : \sigma_a^2 = 0$ используются выражения для математических ожиданий средних квадратов $\frac{Q_1}{r-1}$ и $\frac{Q_2}{n-r}$. Непосредственными вычислениями можно убедиться, что

$$\mathbf{E}\left(\frac{Q_2}{n-r}\right) = \sigma^2, \quad \mathbf{E}\left(\frac{Q_1}{r-1}\right) = \sigma^2 + k\sigma_a^2, \quad \text{где}$$

$$k = \frac{1}{r-1}\left(n - \frac{1}{n}\sum_{i=1}^{r} n_i^2\right). \tag{75}$$

Используем выражение $\bar{x}_i = \mu + a_i + \bar{\epsilon}_i$ для вычисления EQ_2:

$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\mu + a_i + \epsilon_{ij} - \mu - a_i - \bar{\epsilon}_i)^2 =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\epsilon_{ij}^2 - \sum_{i=1}^{r} n_i\bar{\epsilon}_i^2, \quad \sum_{i=1}^{r} n_i = n,$$

$$EQ_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i} E\epsilon_{ij}^2 - \sum_{i=1}^{r} n_i E\bar{\epsilon}_i^2 = n\sigma^2 - \sum_{i=1}^{r} n_i\frac{\sigma^2}{n_i} = (n-r)\sigma^2\,.$$

Для вычисления EQ_1 используем выражение

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{r} n_i\bar{x}_i = \mu + \frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon},$$

$$Q_1 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left((a_i + \bar{\epsilon}_i) - \left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}\right)\right)^2 =$$

$$= \sum_{i=1}^{r} n_i(a_i + \bar{\epsilon}_i)^2 - n\left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}\right)^2;$$

Считаем математическое ожидание:

$$EQ_1 = \sum_{i=1}^{r} n_i E(a_i + \bar{\epsilon}_i)^2 - nE\left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}\right)^2 =$$

$$= \sum_{i=1}^{r} n_i D(a_i + \bar{\epsilon}_i) - nD\left(\frac{1}{n}\sum_{i=1}^{r} n_i a_i + \bar{\epsilon}_i\right) =$$

$$= \sum_{i=1}^{r} n_i\left(\sigma_a^2 + \frac{\sigma^2}{n_i}\right) - n\left(\sigma_a^2\frac{1}{n^2}\sum_{i=1}^{r} n_i^2 + \frac{\sigma^2}{n}\right) =$$

$$= (r-1)\sigma^2 + \sigma_a^2\left(n - \frac{1}{n}\sum_{i=1}^{r} n_i^2\right) = (r-1)\left(\sigma^2 + k\sigma_a^2\right),$$

где $k = \dfrac{1}{r-1}\left(n - \dfrac{1}{n}\sum_{i=1}^{r} n_i^2\right)$.

Если используется одинаковое число повторений в каждой подпопуляции, то есть $n_1 = \ldots = n_r = m$, то $k = m$. При справедливости гипотезы $H_0 : \sigma_a^2 = 0$ статистики $\frac{Q_2}{n-r}$ и $\frac{Q_1}{r-1}$ имеют одинаковые средние σ^2, поэтому можно построить отношение Фишера

$$F = \frac{n-r}{r-1} \cdot \frac{Q_1}{Q_2} \ \sim \ F(r-1, n-r)\,.$$

Значимость критерия равна $\alpha_* = P\{F > F_*\}$, где F_* наблюдаемое значение статистики. Гипотеза об отсутствии влияния фактора отвергается при $\alpha_* < \alpha = 0.05$. что в случае однофакторного дисперсионного анализа значимости моделей со случайными и фиксированными эффектами совпадают. Рассматривая в качестве k выражение (75), получаем оценку дисперсии σ_a^2:

$$\widehat{\sigma}_a^2 = \frac{1}{k}\left(\frac{Q_1}{r-1} - \frac{Q_2}{n-r}\right).$$

6.10. Двуфакторная модель со случайными эффектами

Модель со случайными эффектами имеет вид:

$$x_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}, \tag{76}$$

где $i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K$, μ - генеральное среднее, случайные дифференциальные эффекты A-фактора a_i независимы и распределены по $\mathcal{N}(0; \sigma_a)$, случайные дифференциальные эффекты B-фактора b_j независимы и распределены по $\mathcal{N}(0; \sigma_b)$, эффекты взаимодействия $(ab)_{ij}$ независимы и распределены по $\mathcal{N}(0; \sigma_{ab})$. Величины e_{ijk} независимы и распределены по $\mathcal{N}(0; \sigma)$. Все величины a_i, b_j, $(ab)_{ij}$, e_{ijk} независимы в сово-

купности.

источник дисперсии	средний квадрат	математическое ожидание
A	MQ_A	$\sigma^2 + K\sigma^2_{ab} + JK\sigma^2_a$
B	MQ_B	$\sigma^2 + K\sigma^2_{ab} + IK\sigma^2_b$
AB	MQ_{AB}	$\sigma^2 + K\sigma^2_{ab}$
R	MQ_R	σ^2

Отсюда получаем выражения для оценок дисперсий:

$$\sigma^2_a = \frac{MQ_A - MQ_{AB}}{JK}, \quad \sigma^2_b = \frac{MQ_B - MQ_{AB}}{IK}, \quad \sigma^2_{ab} = \frac{MQ_{AB} - MQ_R}{K}.$$

Для проверки гипотезы $H_0 : \sigma^2_{ab} = 0$ используется отношение $F = \frac{MQ_{AB}}{MQ_R}$, для $H_0 : \sigma^2_a = 0$ используется отношение $F = \frac{MQ_A}{MQ_{AB}}$, для $H_0 : \sigma^2_b = 0$ отношение $F = \frac{MQ_B}{MQ_{AB}}$ с соответствующими степенями свободы.

ПРИМЕР (продолжение) При проверке гипотезы $H_0 : \sigma^2_{ab} = 0$ об отсутствии взаимодействия получаем тот же критерий, что и в случае модели с фиксированными эффектами. $p = 0.16$, взаимодействие незначимо.

При проверке гипотезы $H_0 : \sigma^2_a = 0$ об отсутствии эффекта условий вскармливания получаем значимость $p = 0.059$, при проверке гипотезы $H_0 : \sigma^2_b = 0$ получаем значимость $p = 0.044$. Значимости отличаются от значимостей в модели с фиксированными эффектами, сохраняется только соотношение между ними: фактор условия содержания более значим, чем фактор вскармливания.

В *StatSoft* наиболее удобной для практического использования дисперсионного анализа с фиксированными и случайными эффектами является программа: *Statistics - Advanced Linear/ Nonlinear Models - Variance Components*. В *Variables* указываем нужные переменные. Результаты счета находятся в таблице, спрятанной под кнопкой *Summary : Components of Variance*.

Средние, соответствующие различным уровням факторов и их пересечениям, можно посмотреть во вкладке *Estimation* за кнопкой *Margimal means*.

6.11. Двуфакторная модель со смешанными эффектами

Модель со смешанными эффектами имеет вид:

$$x_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk}, \tag{77}$$

где $i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K,$

μ - генеральное среднее,

α_i – i-й дифференциальный эффект A-фактора,

b_j – случайные эффекты B-фактора независимы и нормально распределены $\mathcal{N}(0; \sigma_b)$, эффекты взаимодействия

$(\alpha b)_{ij}$ независимы и нормально распределены $\mathcal{N}(0; \sigma_{ab})$.

Величины e_{ijk} независимы распределены по $\mathcal{N}(0; \sigma)$. Все величины b_j, $(\alpha b)_{ij}$, e_{ijk} независимы в совокупности.

источник дисперсии	средний квадрат	математическое ожидание
A	MQ_A	$\sigma^2 + K\sigma^2_{ab} + \frac{JK\sum\limits_{i=1}^{I}\alpha_i^2}{I-1}$
B	MQ_B	$\sigma^2 + K\sigma^2_{ab} + IK\sigma^2_b$
AB	MQ_{AB}	$\sigma^2 + K\sigma^2_{ab}$
R	MQ_R	σ^2

Для проверки гипотезы

$H_0 : \sigma^2_{ab} = 0$ используется отношение $F = \frac{MQ_{AB}}{MQ_R}$, для

$H_0 : \sigma^2_b = 0$ используется отношение $F = \frac{MQ_B}{MQ_{AB}}$, для

H_0 : все $\alpha_i = 0$ отношение $F = \frac{MQ_A}{MQ_{AB}}$ с соответствующими степенями свободы. Заметим, что в случае двуфакторного плана результаты дисперсионного анализа в случае хотя бы одного случайного фактора совпадают.

6.12. Двуфакторная модель с группировкой

ПРИМЕР. Пусть имеется фактор диеты A с $r = 4$ уровнями. Зависимой переменной является дважды ($K = 2$) измеряемое количество выдыхаемого азота x_{ijk}, $i = 1, \ldots, I;\ \ j = 1, \ldots, J,\ k = 1, \ldots, K$. Фактор B является фактором испытуемого, $J = 4$.

Данные для двуфакторного плана с группировкой

	диета 1		диета 2		диета 3		диета 4	
B	1 набл.	2 набл.	1 набл.	2 набл.	1 набл.	2 набл.	1 набл.	2 набл.
1	4.079	4.859	4.368	5.668	4.169	5.709	4.928	5.608
2	3.541	5.047	3.752	5.848	4.416	5.666	4.941	5.291
3	3.298	4.679	3.802	4.844	4.123	5.059	4.674	5.038
4	2.871	4.648	3.578	5.393	4.403	4.496	4.905	5.208

Пусть фактор A является случайным фактором. Тогда модель имеет вид:

$$x_{ijk} = \mu + a_i + b_{j(i)} + e_{ijk},$$

где μ - генеральное среднее, величины a_i независимы и распределены по $\mathcal{N}(0; \sigma_a)$, величины $b_{j(i)}$ - независимы и распределены по $\mathcal{N}(0; \sigma_{b(a)})$. Величины e_{ijk} независимы распределены по $\mathcal{N}(0; \sigma)$. Все величины $a_i, b_{j(i)}, e_{ijk}$ независимы в совокупности. $\hat{b}_{j(i)} = \bar{x}_{ij\cdot} - \bar{x}_{i\cdot}$.

Таблица двуфакторного плана с группировкой

источник дисперсии	сумма квадратов	степени свободы	средний квадрат
фактор A	$Q_A = JK \sum_{i=1}^{I} (\bar{x}_{i\cdot} - \bar{x})^2$	$\nu_A = I - 1$	$MQ_A = \frac{Q_A}{\nu_A}$
фактор B (внутриА)	$Q_{B(A)} = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{x}_{ij\cdot} - \bar{x}_{i\cdot})^2$	$\nu_{B(A)} = I(J-1)$	$MQ_{B(A)} = \frac{Q_{B(A)}}{\nu_{B(A)}}$
остаток (ошибка)	$Q_R = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - \bar{x}_{ij\cdot})^2$	$\nu_R = IJ(K-1)$	$MQ_R = \frac{Q_R}{\nu_R}$
полная	$Q_T = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - \bar{x})^2$	$\nu_T = IJK - 1$	

Смешанный двуфакторный план с группировкой можно записать в виде:

$$x_{ijk} = \mu + \alpha_i + b_{j(i)} + e_{ijk},$$

α_i – дифференциальный эффект, определяемый i-м уровнем фактора A.

Приведем ожидания средних квадратов для двуфакторного плана с группировкой в случае модели со случайными эффектами:

$$\mathbf{E}\left(MQ_A\right) = \sigma^2 + K\sigma^2_{b(a)} + KJ\sigma^2_a,$$
$$\mathbf{E}\left(MQ_{B(A)}\right) = \sigma^2 + K\sigma^2_{b(a)}, \quad \mathbf{E}\left(MQ_R\right) = \sigma^2,$$

и в случае модели со смешанными эффектами:

$$\mathbf{E}\left(MQ_A\right) = \sigma^2 + K\sigma^2_{b(a)} + \frac{JK\sum\limits_{i=1}^{I}\alpha_i^2}{I-1},$$
$$\mathbf{E}\left(MQ_{B(A)}\right) = \sigma^2 + K\sigma^2_{b(a)}, \quad \mathbf{E}\left(MQ_R\right) = \sigma^2.$$

Для проверки гипотез используем отношения

$$H_0 : \sigma^2_{b(a)} = 0 \qquad F = \frac{MQ_{B(A)}}{MQ_R},$$
$$H_0 : \sigma^2_a = 0 \text{ или} H_0 : \text{все } \alpha_i = 0 \quad F = \frac{MQ_A}{MQ_{B(A)}}.$$

В *Statistics - Advanced Linear/ Nonlinear Models - Variance Components*. Фактор A можно считать фиксированным или случайным эффектом (в зависимости от модели). Фактор B считать случайным эффектом. В *Model* выбрать *Hierarchically nested design* (и указать вложенность фактора B кнопкой *codes identify levels within other factors*.

Результаты дисперсионного анализа

источник дисперсии	сумма квадратов	число ст.св.	средний квадрат	F	p
Диета	3.711	3	1.237	8.14	0.0032
Объект(диета)	1.828	12	0.152	0.20	0.9960
Остаток	12.020	16	0.751		
Полная	17.559	31			

6.13. Дисперсионный анализ для зависимых выборок

Рассмотрим данные вида $x_{(i)jk}$, где $i = 1, \ldots, r$ уровни фактора A (пол) с фиксированными эффектами, $j = 1, \ldots, n$ – индивиды (фактор B), $k = 1, \ldots, t$ уровни временного фактора C со случайными эффектами. Поскольку фактор A является следствием фактора индивида, то его индекс при наличии индекса j заключаем в скобки, $n_1 + n_2 + \ldots + n_r = n$. Требуется выяснить влияние факторов A и C на переменную X (количество пройденных метров по ТШХ в три момента времени: при поступлении в стационар, при выписке и через три месяца). Построим модель:

$$x_{(i)jk} = \mu + \alpha_i + c_k + e^1_{(i)j} + (\alpha c)_{ik} + e_{(i)jk},$$

где $e^1_{(i)j}$ – ошибка, обусловленная влиянием индивида, $e_{(i)jk}$ – ошибка наблюдения. Оценкой влияния фактора A является разность

$$\widehat{\alpha} = \bar{x}_{i..} - \bar{x}_{...},$$

где $x_{i..}$ есть среднее по каждому полу, $x_{...}$ – общее среднее. Общий источник вариации Q с числом степеней свободы, равным $\nu = nt - 1$, имеет вид:

$$Q = \sum_{j=1}^{n} \sum_{k=1}^{t} (x_{(i)jk} - \bar{x}_{...})^2 .$$

Источник вариации, обусловленный влиянием фактора A, с числом степеней свободы, равным $\nu_A = r - 1$, имеет вид:

$$Q_A = t \sum_{i=1}^{r} n_i (\bar{x}_{i..} - \bar{x}_{...})^2.$$

Вариация Q_A является частью вариации Q_1, обуславливающей различие индивидов, с числом степеней свободы, равным $\nu_1 = n - 1$:

$$Q_1 = t \sum_{j=1}^{n} (\bar{x}_{(i)j.} - \bar{x}_{...})^2 =$$

$$= t \sum_{i=1}^{r} n_i (\bar{x}_{i..} - \bar{x}_{...})^2 + t \sum_{j=1}^{n} (\bar{x}_{(i)j.} - \bar{x}_{i..})^2 = Q_A + Q_{1e} ,$$

где усредненный показатель по каждому индивиду вычисляется как

$$\bar{x}_{(i)j\cdot} = \frac{1}{t}\sum_{k=1}^{t} x_{(i)jk}\,.$$

Ошибка Q_{1e} имеет число степеней свободы, равное $\nu_{1e} = n - r$. Из оставшегося после исключения влияния индивида источник вариации

$$Q_2 = Q - Q_1 = \sum_{j=1}^{n}\sum_{k=1}^{t}(x_{(i)jk} - \bar{x}_{(i)j\cdot})^2$$

выделим вариацию, обуславливающую различие по времени. Источник вариации, обусловленный влиянием временного фактора C, с числом степеней свободы, равным $\nu_C = t - 1$, имеет вид:

$$Q_C = n\sum_{k=1}^{t}(x_{\cdot\cdot k} - \bar{x}_{\cdot\cdot\cdot})^2\,,$$

а источник вариации, обусловленный влиянием взаимодействия фактора A и фактора времени C, с числом степеней свободы, равным $\nu_{AC} = (t - 1)(r - 1)$, имеет вид:

$$Q_{AC} = \sum_{i=1}^{r} n_i \sum_{k=1}^{t}(\bar{x}_{i\cdot k} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot\cdot k} + \bar{x}_{\cdot\cdot\cdot})^2\,.$$

Ошибка имеет вид:

$$Q_{err} = Q_2 - Q_C - Q_{AC} = \sum_{j=1}^{n}\sum_{k=1}^{t}(x_{(i)jk} - \bar{x}_{(i)j\cdot} - \bar{x}_{i\cdot k} + \bar{x}_{i\cdot\cdot})^2\,.$$

Ее число степеней свободы равно $\nu_{err} = (t - 1)(n - r)$.

Для проверки гипотезы о том, что все дифференциальные эффекты фактора A равны нулю, вычисляется статистика

$$F = \frac{MQ_A}{MQ_{1e}} = \frac{Q_A/\nu_A}{Q_{1e}/\nu_{1e}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_A = r - 1$ и $\nu_{1e} = n - r$.

Для проверки гипотезы о том, что случайные эффекты временного фактора имеют нулевую дисперсию, вычисляется статистика

$$F = \frac{MQ_C}{MQ_{AC}} = \frac{Q_C/\nu_C}{Q_{AC}/\nu_{AC}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_C = t - 1$ и $\nu_{AC} = (t-1)(r-1)$.

В случае предположения фиксированных эффектов временного фактора C вычисляется статистика

$$F = \frac{MQ_C}{MQ_{err}} = \frac{Q_C/\nu_C}{Q_{err}/\nu_{err}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_C = t - 1$ и $\nu_{err} = (t-1)(n-r)$.

Для проверки эффекта взаимодействия вычисляется статистика

$$F = \frac{MQ_{AC}}{MQ_{err}} = \frac{Q_{AC}/\nu_{AC}}{Q_{err}/\nu_{err}},$$

которая в случае справедливости нулевой гипотезы имеет распределение Фишера с числом степеней свободы $\nu_{AC} = (t-1)(r-1)$ и $\nu_{err} = (t-1)(n-r)$.

Для обработки в $StatSoft$ данных о ТХШ, дифференцируемых по признаку пол, в трех временных точках нужно представить исходные данные в виде четырех столбцов: sex, $m1$, $m2$ и $m3$. В *Statistics* выбираем $ANOVA$ - *Repeated measures* $ANOVA$. В $Variables$ выбираем переменные: в *Dependent variable list* отмечаем $m1, m2, m3$, в *Categorial predictors* ($factors$) отмечаем sex. Обязательно нажать кнопку $Within\ effects$, включением которой данные о ТХШ в трех временных точках собираются в одну переменную, обычно ее называют $R1$. В противном случае, вместо нужной процедуры пойдет многомерный дисперсионный анализ. Кроме того, для получения указанных ранее сумм квадратов, в опциях $Sum\ of\ squares$ нужно указать последовательный тип $Type\ 1\ sequental$.

Исходные данные: $n = 37$, $r = 2$, $t = 3$. В результате счета получаем, что $Q_A = 173604$, $\nu_A = 1$, $Q_{1e} = 721179$, $\nu_{1e} = 35$, для проверки значимости фактора пола используем статистику $F = \frac{173604/1}{721179/35} = 8.42$, $p = 0.006364$, по которой видно, что различие в средних 310 метров у женщин и 406 метров у мужчин является значимым.

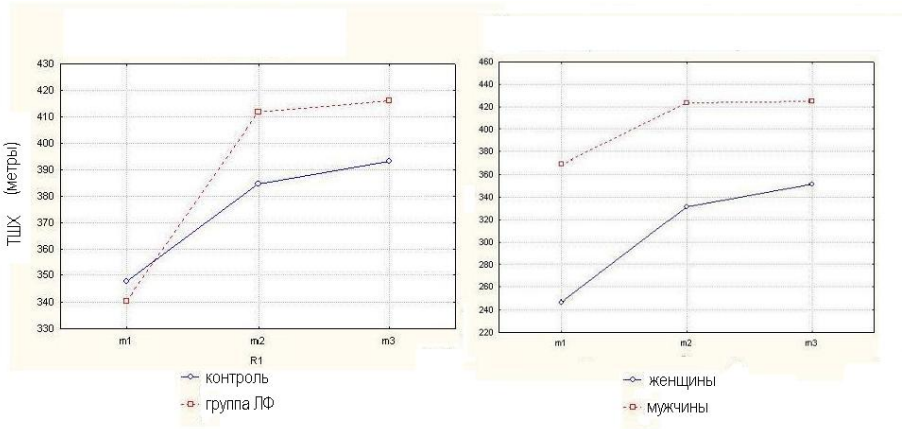
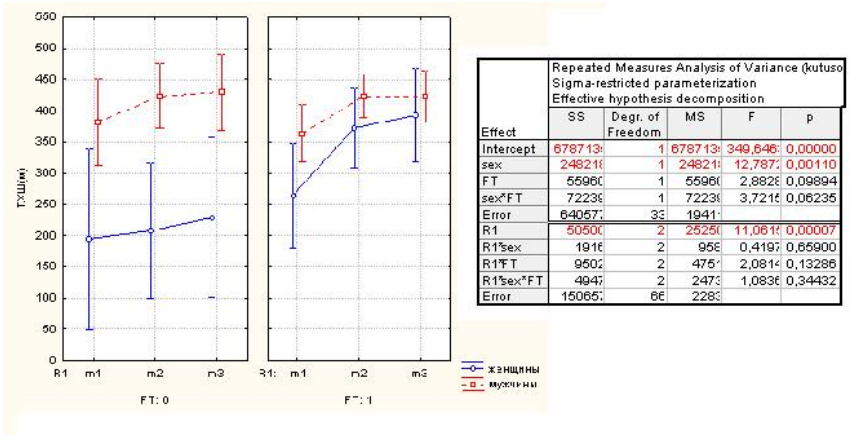


Рис. 4. Эффект взаимодействия (по полу $p = 0.199$ и по ЛФ $p = 0.32$).



Repeated Measures Analysis of Variance (kutuso
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	678713	1	678713	349,646	0,00000
sex	24821	1	24821	12,787	0,00110
FT	5596	1	5596	2,8828	0,09894
sex*FT	7223	1	7223	3,7215	0,06235
Error	64057	33	1941		
R1	5050	2	2525	11,061	0,00007
R1*sex	191	2	95	0,4197	0,65900
R1*FT	950	2	475	2,0814	0,13286
R1*sex*FT	494	2	247	1,0836	0,34432
Error	15065	66	228		

Рис. 5. Динамика ТХШ в зависимости от пола и лечебной физкультуры.

Если эффекты времени фиксированы, то влияние времени значимо:

$$Q_C = 100805, \nu_C = (r-1)(t-1) = 2,$$
$$Q_{err} = 160441,\ \nu_{err} = (t-1)(n-r) = 2 \cdot (37-2) = 70.$$

$F = 21.99$, $p < 0.0001$. Если эффекты считать случайными, то $p = 0.06$. Графики средних указывают на значимое отличие пройденных метров по ТХШ при поступлении в стационар и при выписке. Эффект взаимодействия не значим, $p = 0.199$.

Аналогичный счет при исследовании влияния фактора лечебной физкультуры на динамику ТХШ показал, что влияние ЛФК не значимо ($p = 0.67$), влияние взаимодействия также не значимо ($p = 0.32$). Значимо влияние фактора времени ($p < 0.001$).

6.14. Поправки Бонферони

В методе LSD (the least significant difference — наименее значимое различие) применяются стандартные t-критерии ко всем возможным парам средних групп. Пусть имеется r выборок $\mathcal{N}(\mu_i, \sigma)$ x_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$, $n = \sum\limits_{i=1}^{r} n_i$, ранее было показано, что

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n-r), \quad \text{где } s^2 = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - x_{..})^2$$

Поправки Бонферрони применяются для того, чтобы частота ложноположительных результатов с поправкой на эффект множественных сравнений не превышала заданное значение.

Пусть имеются нулевые гипотезы $\mathcal{H}_1, \ldots, \mathcal{H}_m$ с соответствующими p-значениями: $p_1, \ldots, p_m$. Например, $m = C_r^2$, если проверяются гипотезы о равенстве всех попарных средних. Определим как вероятность совершить хотя бы одну ошибку в выводе из m выводов. Если осуществляется n независимых испытаний с вероятностью ошибки α в каждом, то $\tilde{\alpha} = 1-(1-\alpha)^n$. В случае зависимых испытаний $\tilde{\alpha} \leqslant 1 - (1-\alpha)^n$. Заметим, что по неравенству Буля

$$\tilde{\alpha} = 1 - (1 - n\alpha + C_n^2\alpha^2 - \ldots) \leqslant n \cdot \alpha.$$

Таким образом, если мы хотим, чтобы вероятность наличия хотя бы одного неверного вывода из m была равна 0.05, то достаточно установить вероятность неверного отклонения нулевой гипотезы равной $0.05/m$ для каждого вывода. Такой метод называется поправкой Бонферрони (англ.). Если гипотезы имеют различные доверительные уровни вероятностей, то достаточно отвергнуть гипотезы, имеющие $\alpha < \tilde{\alpha}/m$.

6.15. Метод множественных сравнений

Определение 2. Сравнением параметров $\beta_1, \ldots, \beta_p$ называется линейная функция $\sum\limits_{i=1}^{p} c_i\beta_i$, где $\sum\limits_{i=1}^{p} c_i = 0$.

Например, имеются три градации категориального признака, из которых две подвергнуты некоторому условию, а одна нет. Для проверки этого утверждения можно рассмотреть разность

$$\psi = \frac{\beta_1 + \beta_2}{2} - \beta_3 \,.$$

Обозначим через $\hat{\psi} = \sum\limits_{i=1}^{p} c_i\hat{\beta}_i$ несмещенную оценку сравнения. В матричном виде

$$\psi_{q,1} = C_{q,p}\beta_{p,1}, \quad rank(C) = q,$$
$$\hat{\psi} = C\hat{\beta} = C(X^TX)^{-}X^TY = AY.$$

Ковариационная матрица оценок находится по формуле

$$\Gamma_\psi = \sigma^2 AA^T, B = AA^T,$$

несмещенной оценкой σ^2 является средний квадрат ошибок

$$s^2 = \frac{R_0^2}{n-r}$$

Теорема 3. *(Метод множественных сравнений Шеффе) Если* $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, $rank(X) = r$, *то случайная величина* $\hat{\psi} \sim \mathcal{N}(\psi, \Gamma_\psi)$ *и не зависит от* $R_0^2/\sigma^2 \sim \chi^{\mathbf{2}}(n-r)$. *Поэтому*

$$\frac{(\hat{\psi}-\psi)^T B^{-1}(\hat{\psi}-\psi)}{qs^2} \sim F(q, n-r).$$

Пусть элементы выборки $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma)$ упорядочены в виде $x^{(1)} < \ldots < x^{(n)}$, разность $R = x^{(n)} - x^{(1)}$ называется размахом выборки.

Пусть s^2 является независимой среднеквадратичной оценкой σ^2 с ν степенями свободы. Таким образом $\nu s^2/\sigma^2 = \chi^2_\nu$ не зависит от R. Случайную величину $R/s = q_{n,\nu}$ называют стьюдентизированным размахом.

Метод Тьюки можно применять для получения совместных доверительных утверждений о сравнениях множества параметров $\theta_1, \ldots, \theta_k$ в терминах несмещенных оценок. Ограничением метода Тьюки является требование одинаковых дисперсий для оценок θ_i. Поэтому если нужна классификация по одному признаку, то объемы должны быть равными.

Теорема 4. *(Метод множественных сравнений Тьюки) Пусть* $\hat{\theta}_1, \ldots, \hat{\theta}_k$ *независимы,* $\mathcal{N}(\theta_i, a^2\sigma^2)$, $\frac{\nu s^2}{\sigma^2} = \chi^2_\nu$ *не зависит от* $\{\theta_i\}$, $\hat{\psi} = \sum_{i=1}^{k} c_i\hat{\theta}_i$, $T = aq_{\alpha,k,\nu}$, *где* $q_{\alpha,k,\nu}$ *верхний* α *предел стьюдентизированного размаха. Тогда вероятность того, что все* $\frac{k(k-1)}{2}$ *разностей* $\{\theta_i - \theta_j\}$ *одновременно удовлетворяют неравенствам*

$$\hat{\theta}_i - \hat{\theta}_j - Ts \le \theta_i - \theta_j \le \hat{\theta}_i - \hat{\theta}_j + Ts$$

равна $1 - \alpha$.

Кроме того, существует обобщение для $\hat{\theta}_1, \ldots, \hat{\theta}_k$, не обязательно независимых, но имеющих одинаковые ковариации и дисперсии.

В критерии Пиллая используется верхний α предел стьюдентизированного максимума модулей $M = max|x_i|/s$. Увеличенным размахом R' называется $max(R, M)$.

В тех случаях, когда главный интерес представляют все разности, причем никакой из них не отдается предпочтения, метод Тьюки дает более узкие интервалы, но применим только в случае одинаковых дисперсий. Преимущество метода Шеффе в том, что он не чувствителен к нарушению о предположении нормальности и равенства дисперсий.

7. Дискриминантный анализ

7.1. Классификация индивидов в случае двух популяций

Пусть имеется две нормально распределенные популяции W_1 и W_2 с одинаковой известной ковариационной матрицей Σ и векторами средних μ_1 и μ_2. Задача дискриминантного анализа состоит в том, чтобы решить, к какой из этих популяций отнести вектор наблюдений $X = (x_1, \ldots, x_p)^T$.

ПРИМЕР. W_1 - больные бронхитом, W_2 - больные пневмонией, x_1 серосодержащие белки в восстановительной форме ($SH1$), x_2 скорость оседания эритроцитов ($SOE1$) в начале заболевания. Средние $SH1 = 7.1 \pm 0.7$ в W_1 и $SH1 = 5.8 \pm 0.8$ в W_2 $SOE1 = 32.9 \pm 4.4$ в W_1 и $SOE1 = 46.0 \pm 5.4$ в W_2. Пусть у некоторого больного $SH1 = 5.5$, $SOE1 = 40$. Можно ли сказать, какая форма заболевания органов дыхания у этого больного?

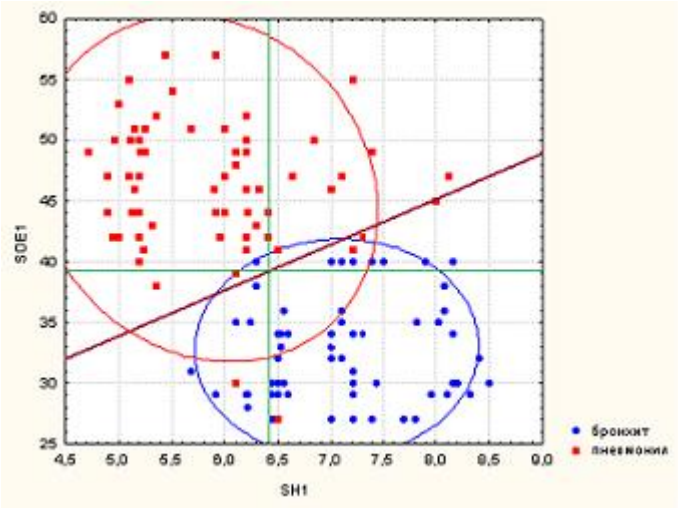


Рис. 6. Классифицирующая прямая $SOE1 = 3.7SH1 + 15.5$.

Ищется так называемая *дискриминантная функция* вида

$$z = \alpha_1 x_1 + \ldots + \alpha_p x_p = \alpha^T X$$

где $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ – некоторый вектор констант, и найти граничное значение c, так что если $z \geq c$, то наблюдение X относится к популяции W_1, и иначе в противоположном случае. Например, $z = -1.97x_1 + 0.53x_2$, $c = 8.22$ (рис.7), отсюда $x_2 = 40 > 36 = 3.7x_1 + 15.5$ при $x_1 = 5.5$, - это область популяции W_2, следовательно, скорее всего у этого больного не бронхит, а пневмония.

Расстояние Махаланобиса

Основная проблема заключается в выборе таких коэффициентов дискриминантной функции, чтобы ошибка классификации была как можно меньше.

Предложение. *Если наблюдение поступило из нормально распределенной популяции* $W \sim \mathcal{N}(\mu, \Sigma)$, *то* $z = \alpha^T X \sim \mathcal{N}(\zeta, \sigma)$, *где* $\zeta = \alpha^T \mu$, $\sigma^2 = \alpha^T \Sigma \alpha$.

Доказательство.

$$1. Ez = E\alpha^T X = \alpha^T EX = \alpha^T \mu,$$

$$2. \sigma^2 = E(\alpha^T X - \alpha^T \mu)^2 = E(\alpha^T (X - \mu))^2 =$$

$$= E(\alpha^T (x - \mu))(\alpha^T (X - \mu))^T = E\alpha^T (X - \mu)(X - \mu)^T \alpha =$$

$$= \alpha^T E(X - \mu)(X - \mu)^T \alpha = \alpha^T \Sigma \alpha.$$

Обозначим через $\zeta_i = \alpha^T \mu_i$ средние значения дискриминантной функции $z = \alpha^T X$ в случае популяции W_i, $i = 1, 2$. Имеет смысл выбрать такой вектор α, чтобы ζ_1 и ζ_2 были как можно дальше удалены друг от друга. Для этого введем расстояние Махаланобиса

$$\Delta^2 = \frac{(\zeta_1 - \zeta_2)^2}{\sigma^2} = \frac{(\alpha^T (\mu_1 - \mu_2))^2}{\alpha^T \Sigma \alpha}.$$

Оно должно быть максимальным. В 1936 году Фишер показал, что *максимальное расстояние Махаланобиса доставляет такой вектор* α, *который*

является решением системы уравнений

$$\Sigma\alpha = \mu_1 - \mu_2. \tag{78}$$

Рассуждения, подтверждающие это, таковы. Обозначим $\mu_1 - \mu_2 = m$. Максимальное расстояние Махаланобиса доставляет вектор $\alpha = \Sigma^{-1}m$. Нетрудно убедиться, что $\Delta^2_{max} = \alpha^T\Sigma\alpha = m^T\Sigma^{-1}m$. Нужно показать, что расстояние Махаланобиса $\Delta^2 = \frac{(\alpha^T m)^2}{\alpha^T\Sigma\alpha}$ для любого другого вектора α будет меньше $\Delta^2_{max} = m^T\Sigma^{-1}m$. Действительно, пусть α является решением другого уравнения $\Sigma\alpha = y$. Следовательно, $\alpha = \Sigma^{-1}y, \alpha^T = y^T\Sigma^{-1}$. Тогда выражение $\Delta^2 \leq \Delta^2_{max}$ равносильно выражению

$$(y^T\Sigma^{-1}m)^2 \leq ((y^T\Sigma^{-1})\Sigma(\Sigma^{-1}y))(m^T\Sigma^{-1}m).$$

То есть справедливо неравенство Коши-Буняковского

$$(y^T\Sigma^{-1}m)^2 \leq (y^T\Sigma^{-1}y))(m^T\Sigma^{-1}m).$$

Заметим, что при справедливости (78) имеет место

$$\sigma^2 = \alpha^T\Sigma\alpha = \alpha^T(\mu_1 - \mu_2) = \zeta_1 - \zeta_2. \tag{79}$$

С другой стороны, $\Delta = \sigma$, так как

$$\Delta^2 = \frac{(\zeta_1 - \zeta_2)^2}{\zeta_1 - \zeta_2} = |\zeta_1 - \zeta_2|,$$

т.е. есть растояние Махаланобиса равно абсолюной величине разности между средними значениями дискриминантной функции.

Вычисление вероятности ошибочной классификации

Если $X \in W_2$, но $z \geq c$, то имеет место ошибка классификации $P(1|2)$, если $X \in W_1$, но $z \leq c$, то имеет место ошибка классификации $P(2|1)$.

Нужно, чтобы *ошибка классификации была минимальной*. Это достигается выбором постоянной, равноудаленной от средних

$$c = \frac{\zeta_1 + \zeta_2}{2}.$$

Покажем это. Пусть граница дискриминантной функции равна $c_\delta = c + \delta$. Вероятность $P_\delta(2|1)$ отнесения наблюдения к популяции W_2, когда оно принадлежит популяции W_1, определяется функцией $\Phi(x)$ стандартного нормального распределения

$$P_\delta(2|1) = P\{z < c + \delta | W_1\} = \Phi\left(\frac{c + \delta - \zeta_1}{\sigma}\right)$$

Учитывая, что $\Delta^2 = \sigma^2 = \zeta_1 - \zeta_2$ и подставляя значение c, имеем

$$P_\delta(2|1) = \Phi\left(\frac{\frac{\zeta_1+\zeta_2}{2} + \delta - \zeta_1}{\Delta}\right) = \Phi\left(-\frac{\Delta}{2} + \frac{\delta}{\Delta}\right).$$

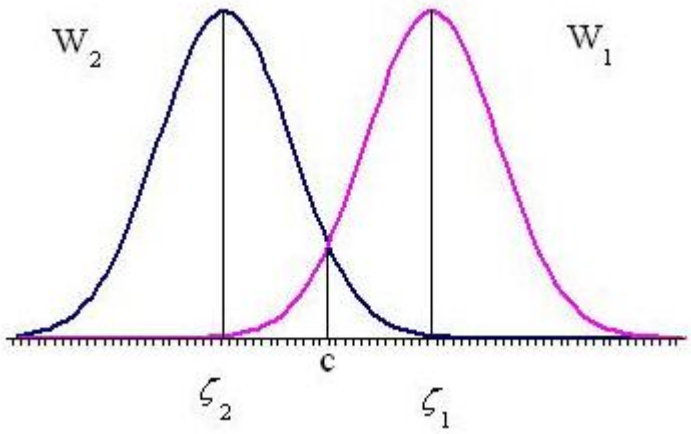


Рис. 7. Распределения дискриминантной функции в случае разных популяций.

Аналогично получаем

$$P_\delta(1|2) = P\{z > c + \delta | W_2\} = 1 - \Phi\left(\frac{c + \delta - \zeta_2}{\sigma}\right) = \Phi\left(-\frac{\Delta}{2} - \frac{\delta}{\Delta}\right).$$

Минимальная вероятность ошибочной классификации

$$P_\delta = (P_\delta(2|1) + P_\delta(1|2))/2$$

достигается при $\delta = 0$ и равна $\Phi(-\frac{\Delta}{2})$. Действительно,

$$2P_\delta = P_\delta(2|1) + P_\delta(1|2) = \Phi\left(-\frac{\Delta}{2} + \frac{\delta}{\Delta}\right) + \Phi\left(-\frac{\Delta}{2} - \frac{\delta}{\Delta}\right) =$$
$$= \Phi\left(-\frac{\Delta}{2}\right) + P\left\{-\frac{\Delta}{2} < z < -\frac{\Delta}{2} + \frac{\delta}{\Delta}\right\} +$$
$$+\Phi\left(-\frac{\Delta}{2}\right) - P\left\{-\frac{\Delta}{2} - \frac{\delta}{\Delta} < z < -\frac{\Delta}{2}\right\} > 2\Phi\left(-\frac{\Delta}{2}\right),$$

так как $P\{-\frac{\Delta}{2} - \frac{\delta}{\Delta} < z < -\frac{\Delta}{2}\} < P\{-\frac{\Delta}{2} < z < -\frac{\Delta}{2} + \frac{\delta}{\Delta}\}$ вследствие того, что на интервале $(-\infty; 0)$ плотность стандартного нормального распределения является возрастающей функцией, а также $\zeta_1 > \zeta_2$ и $-\frac{\Delta}{2} < 0$.

Такая процедура классификации, не зависящая от априорной вероятности i-ой популяции, называется эвристической. В случае, когда априорные вероятности не одинаковы, требуется другой подход, связанный с байесовской процедурой классификации.

Байесовская процедура классификации

Определим априорные вероятности q_i того, что объект принадлежит популяции W_i, в частности, $q_1 = P(W_1), q_2 = P(W_2)$ $(q_1 + q_2 = 1)$, апостериорные вероятности $P(W_1|X)$ и $P(W_2|X)$, определяемые как вероятности того, что объект принадлежит к данной популяции при данном векторе наблюдений X, а также $P(X|W_i)$ - условную вероятность получения вектора наблюдений X, если известно, что объект относится к популяции W_i.

Байесовская процедура классификации основана на соотношениями между апостериорными вероятностями: если $P(W_1|X) \geq P(W_2|X)$, то объект относится к популяции W_1; если $P(W_1|X) \leq P(W_2|X)$, то объект относится к популяции W_2. Апостериорные вероятности могут быть вычислены

по формуле Байеса

$$P(W_i|X) = \frac{P(W_i)P(X|W_i)}{P(W_1)P(X|W_1) + P(W_2)P(X|W_2)}.$$

Вероятность $P(X|W_i)$ получения наблюдения X, если известно, что объект принадлежит популяции W_i, может быть заменена на плотности $f_i(x)$:

$$P(W_i|X) = \frac{q_i f_i(x)}{q_1 f_1(x) + q_2 f_2(x)}.$$

Итак, байесовская процедура заключается в следующем: если

$$P(W_1|X) > P(W_2|X) \iff \frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 1,$$

тогда наблюдение X относится к популяции W_1, в противном случае к W_2. В работах Рао (1965) и Андерсон (1958) показано, что эта процедура минимизирует вероятность ошибочной классификации, вычисляемую по формуле полной вероятности,

$$q_1 P(2|1) + q_2 P(1|2).$$

Для нормально распределенных популяций с векторами средних μ_1 и μ_2 байесовская процедура классификации эквивалентна

$$ln\frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 0 \Longleftrightarrow ln\frac{q_1}{q_2} + lnf_1(x) - lnf_2(x) \geq 0.$$

Используем формулу плотности многомерного нормального распределения

$$\begin{aligned} lnf_1(x) - lnf_2(x) = ln\left\{(2\pi)^{-p/2}|\Sigma|^{-1/2}\right\} - \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \\ -ln\left\{(2\pi)^{-p/2}|\Sigma|^{-1/2}\right\} + \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) = \\ = -\frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}\mu_1^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \\ +\frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu_2^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 = \\ = \frac{1}{2}(\mu_1^T - \mu_2^T)\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2\Sigma^{-1}\mu_2 = \end{aligned}$$

Поскольку первые два слагаемые одинаковы, их можно объединить, кроме того, прибавим и отнимем одно и то же слагаемое $\mu_1^T\Sigma^{-1}\mu_2/2$, получаем с учетом (78) и определения $\zeta_i = \alpha^T\mu_i$

$$= x^T\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}\mu_1^T\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}\mu_2 =$$
$$= x^T\alpha - \frac{1}{2}\mu_1^T\alpha - \frac{1}{2}\alpha^T\mu_2 = x^T\alpha - \frac{\zeta_1+\zeta_2}{2}.$$

То есть байесовская процедура классификации для нормально распределенных популяций определяется неравенством

$$x^T\alpha \geq \frac{\zeta_1+\zeta_2}{2} + ln\frac{q_2}{q_1}.$$

Усовершенствование процедуры классификации заключается во включении в рассмотрение стоимости ошибочной классификации. Обозначим через $C(1|2)$ и $C(2|1)$ стоимости потерь при классификации популяции W_1, когда на самом деле популяция W_2 и наоборот. Обобщенная процедура классификации основана на неравенстве

$$x^T\alpha \geq \frac{\zeta_1+\zeta_2}{2} + ln\frac{q_2C(1|2)}{q_1C(2|1)}.$$

При выполнении этого неравенства объект относится к популяции W_1. Эта процедура минимизирует ожидаемую стоимость ошибочной классификации $q_1C(2|1)P(2|1) + q_2C(1|2)P(1|2)$. Вероятность ошибочной классификации определяется следующим образом. Пусть

$$K = ln\frac{q_2C(1|2)}{q_1C(2|1)},\ T = \frac{\zeta_1+\zeta_2}{2} + K.\ \text{Тогда}$$
$$P(2|1) = \Phi\left(\frac{T-\zeta_1}{\sigma}\right) = \Phi\left(\frac{\frac{\zeta_1+\zeta_2}{2}+K-\zeta_1}{\Delta}\right) =$$
$$= \Phi\left(\frac{\frac{\zeta_2-\zeta_1}{2}+K}{\Delta}\right) = \Phi\left(\frac{-\frac{\Delta^2}{2}+K}{\Delta}\right) = \Phi\left(-\frac{\Delta}{2}+\frac{K}{\Delta}\right),$$

так как $\sigma = \Delta$ и $\Delta^2 = \zeta_1 - \zeta_2$, через $\Phi(x)$ обозначена функция стандартного нормального распределения. Аналогично, используя $\Phi(-x) = 1 - \Phi(x)$,

получаем

$$P(1|2) = 1 - \Phi\left(\frac{T-\zeta_2}{\sigma}\right) = 1 - \Phi\left(\frac{\Delta}{2} + \frac{K}{\Delta}\right) = \Phi\left(-\frac{\Delta}{2} - \frac{K}{\Delta}\right).$$

Процедура классификации в случае двух многомерных нормальных популяций при неизвестных параметрах

При неизвестных параметрах распределения можно воспользоваться выборочными оценками. Обобзначим p-мерный вектор, соответствующий значениям p признаков для j-го индивида из i-й популяции через X_j^i. Рассмотрим две независимые выборки: $X_1^1, \ldots, X_{n_1}^1$ объема n_1 из популяции W_1, другая $X_1^2, \ldots, X_{n_1}^2$ объема n_2 из популяции W_2. Выборочные оценки среднего и ковариационной матрицы для популяции W_i имеют вид

$$\bar{X}^i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_j^i,\ i = 1, 2,$$

$$S_i = \left\{s_{kj}^i\right\}_{k,j=1}^p = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_j^i - \bar{X}^i)(X_j^i - \bar{X}^i)^T.$$

В качестве объединенной выборочной ковариационной матрицы используем выражение:

$$S = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)S_1 + (n_2 - 1)S_2).$$

Для построения дискриминантной функции решается система уравнений $\Sigma\alpha = \mu_1 - \mu_2$, где μ_i заменяется на $\bar{X}^i$, а матрица Σ оценкой S. Полученные оценки a используются для построения дискриминантной функции $z = a^T X = \sum_{j=1}^{p} a_j x_j$. Граничное значение дискриминантной функции определяется в виде

$$z_k = \frac{\bar{z}_1 + \bar{z}_2}{2} + \ln\frac{q_2 C(1|2)}{q_2 C(2|1)},$$

где $\bar{z}_i = a^T \bar{X}_i$ среднее значение дискриминантной функции, если X относится к популяции W_i. Выборочная оценка s_z^2 дисперсии дискриминантной

функции z имеет вид $s_z^2 = a^T S a$. Выборочное расстояние Махаланобиса

$$D^2 = \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2}$$

является оценкой Δ^2. Априорная вероятность оценивается как $q_1 = \frac{n_1}{n_1+n_2}$, $q_2 = \frac{n_2}{n_1+n_2}$. Нулевая гипотеза вида $H_0 : \Delta^2 = 0$ эквивалентна $H_0 : \mu_1 = \mu_2$ и проверяется на основе статистики

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \cdot \frac{n_1 n_2}{n_1 + n_2} D^2$$

которая имеет в случае справедливости нулевой гипотезы распределение Фишера $\mathcal{F}(p, n_1+n_2-p-1)$. Вероятность ошибочной классификации оценивается по формулам: $P(2|1) = \frac{m_1}{n_1}, \quad P(1|2) = \frac{m_2}{n_2}$, где m_1 число индивидов из популяции W_1, которые на основе процедуры классификации относятся к популяции W_2; m_2 определяется аналогично.

7.2. Классификация в случае k популяций

Пусть $\mathbf{x} = (x_1, \ldots, x_p)^T$ некоторый вектор наблюдений. Задача состоит в том, чтобы отнести его к одной из популяций $W_1, \ldots, W_k$ с произвольным распределением. Обозначим через $q_i = P(W_i)$ априорную вероятность того, что случайный вектор $\mathbf{X} \in W_i$, $f_i(\mathbf{x})$ – плотность распределения популяции W_i, C_{ij} – стоимость отнесения к популяции W_i элемента из популяции W_j, P_{ij} – вероятность отнесения к популяции W_i элемента из популяции W_j. Будем считать, что все параметры известны. Апостериорная вероятность $P(W_j|\mathbf{x})$ выражается по формуле Байеса:

$$P(W_j|\mathbf{x}) = \frac{q_j f_j(\mathbf{x})}{\sum_{i=1}^{k} q_i f_i(\mathbf{x})}. \tag{80}$$

Условная потеря от отнесения $\mathbf{x}$ к популяции W_t выразится в виде

$$L_t = \sum_{j=1}^{k} C_{tj} P(W_j|\mathbf{x}) = \frac{\sum\limits_{j=1}^{k} C_{tj} q_j f_j(\mathbf{x})}{\sum\limits_{i=1}^{k} q_i f_i(\mathbf{x})}.$$

Минимальность потери эквивалентна максимальности дискриминанта S_t, равного по определению $S_t \stackrel{df}{=} -\sum\limits_{j=1}^{k} C_{tj} q_j f_j(\mathbf{x})$. В случае, когда стоимость ошибочной классификации не имеет значения, то есть $C_{tj} = 1,\ t \neq j,\ C_{tt} = 0$, максимизация дискриминанта

$$S_t = -\sum_{j=1}^{k} q_j f_j(\mathbf{x})|_{t \neq j} = c + q_t f_t(\mathbf{x}), \quad \text{где} \ \ c = -\sum_{j=1}^{k} q_j f_j(\mathbf{x}), \tag{81}$$

эквивалентна максимизации числителя апостериорной вероятности (80).

Таким образом, обобщенная байесовская процедура относит наблюдение $\mathbf{x}$ к W_i, если дискриминант S_k является максимальным. При этом безусловная ожидаемая стоимость ошибочной классификации минимизируется. Покажем, что она имеет вид

$$M = \sum_{j=1}^{k} q_j (\sum_{t=1}^{k} P_{tj} C_{tj}) \tag{82}$$

Действительно, она вычисляется как

$$M = \sum_{t=1}^{k} M_t,$$

где $M_t = q_1 P_{t1} C_{t1} + \ldots + q_k P_{tk} C_{tk}$ есть ожидаемая стоимость ошибочной классификации при отнесении элемента к популяции W_t. Отсюда

$$M = q_1 \sum_{t=1}^{k} P_{t1} C_{t1} + \ldots + q_k \sum_{t=1}^{k} P_{tk} C_{tk} = \sum_{j=1}^{k} q_j \sum_{t=1}^{k} P_{tj} C_{tj}.$$

Классификация k нормально распределенных популяций

Пусть популяция W_i имеет нормальное распределение $N(\mu_\mathbf{i}, \mathbf{\Sigma})$, а стоимости ошибочной классификации не имеют значения: $C_{ij} = 1$ при $i \neq j$ и $C_{ii} = 0$. Введем в рассмотрение дискриминантную функцию вида

$$\delta_i = \ln(q_i f_i(\mathbf{x})) - C_1, \quad \text{где} \quad C_1 = -\frac{p}{2} ln(2\pi) - \frac{1}{2} ln|\mathbf{\Sigma}| + \frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}^{-1}\mathbf{x} \qquad (83)$$

постоянная, которая оказывается одинаковой для всех популяций W_i.

$$\delta_i = \ln(q_i f_i(\mathbf{x})) - C_1 = \ln(q_i (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}}) exp\{-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{i})\} - C_1 =$$

$$= \ln q_i - \frac{p}{2} ln(2\pi) - \frac{1}{2} ln|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{i}) - C_1 =$$

$$= \ln q_i + {\mu_\mathbf{i}}^T \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}{\mu_\mathbf{i}}^T \mathbf{\Sigma}^{-1}\mu_\mathbf{i} = \alpha_i^T \mathbf{x} + \gamma_i + \ln q_i,$$

где $\alpha_i = \mathbf{\Sigma}^{-1}\mu_\mathbf{i}$, $\gamma_i = -\frac{1}{2}{\mu_\mathbf{i}}^T \mathbf{\Sigma}^{-1}\mu_\mathbf{i}$. Вектор $\mathbf{x}$ относится к популяции с максимальным значением дискриминантной функции.

Апостериорная вероятность

Апостериорная вероятность имеет вид

$$P(W_i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{\sum\limits_{j=1}^{k} q_j f_j(\mathbf{x})} = \frac{e^{\delta_i}}{\sum\limits_{j=1}^{k} e^{\delta_j}},$$

т. к. из (83) для выражения постоянной C_1 получаем

$$\delta_i = \ln q_i f_i(\mathbf{x}) - C_1 = \ln q_i f_i(\mathbf{x}) - \ln e^{C_1} = \ln \frac{q_i f_i(\mathbf{x})}{e^{C_1}},$$

$$q_i f_i(x) = e^{C_1} e^{\delta_i}, \quad \text{то}$$

$$P(W_i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{\sum\limits_{j=1}^{k} q_j f_j(\mathbf{x})} = \frac{e^{C_1} e^{\delta_i}}{\sum\limits_{j=1}^{k} e^{C_1} e^{\delta_j}} = \frac{e^{\delta_i}}{\sum\limits_{j=1}^{k} e^{\delta_j}}.$$

В случае неизвестных параметров μ и $\mathbf{\Sigma}$ можно воспользоваться их выборочными оценками. Пусть n_i объем выборки, относящейся к популяции W_i, $\bar{\mathbf{x}}_i$ вектор средних используется в качестве оценки $\mu_\mathbf{i}$, S_i – оценка

ковариационной матрицы для популяции W_i. В качестве оценки общей ковариационной матрицы берется оценка

$$S = \frac{\sum_{i=1}^{k}(n_i - 1)S_i}{\sum_{i=1}^{k} n_i - k}.$$

Тогда в качестве оценки дискриминантных функций рассматриваются

$$d_i = a^T\mathbf{x} + g_i + \ln q_i,$$

где $a_i = S^{-1}\bar{\mathbf{x}}_i,\ g_i = -\frac{1}{2}\bar{\mathbf{x}}_i^T S^{-1}\bar{\mathbf{x}}_i$. Вектор $\mathbf{x} \in W_i$, если d_i максимально.

Если $k = 2$, то $\delta_1 = \alpha_1^T\mathbf{x} + \gamma_1 + \ln q_1$, $\delta_2 = \alpha_2^T\mathbf{x} + \gamma_2 + \ln q_2$. Наблюдение $\mathbf{x}$ относится к популяции W_i, если $\delta_1 \geq \delta_2$, что эквивалентно

$$\delta_1 - \delta_2 \geq 0 \iff (\alpha_1^T - \alpha_2^T)\mathbf{x} + (\gamma_1 - \gamma_2) + \ln\frac{q_1}{q_2} \geq 0.$$

Обозначим через $\alpha = \alpha_1 - \alpha_2 = \mathbf{\Sigma}^{-1}(\mu_\mathbf{1} - \mu_\mathbf{2})$.

$$\gamma_1 - \gamma_2 = -\frac{1}{2}\mu_\mathbf{2}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} + \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{1} + \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} - \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} =$$

$$\frac{1}{2}(\mu_\mathbf{1} - \mu_\mathbf{2})^T\mathbf{\Sigma}^{-1}\mu_\mathbf{2} + \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}^{-1}(\mu_\mathbf{1} - \mu_\mathbf{2}) = \frac{1}{2}\alpha^T\mu_\mathbf{2} + \frac{1}{2}\mu_\mathbf{1}{}^T\alpha = \frac{\zeta_1 + \zeta_2}{2}.$$

Таким образом получаем, что общая процедура классификации в случае $k = 2$ эквивалентна байесовской процедуре классификации вида

$$\alpha^T\mathbf{x} \geq= \frac{\zeta_1 + \zeta_2}{2} + \ln\frac{q_2}{q_1}.$$

ПРИМЕР (продолжение). Матрица коэффициентов дискриминантных функций имеет вид:

	W_1	W_2
$SH1$	12.81	10.84
$SOE1$	1.46	1.99
$Constant$	−69.95	−78.15

получаем выражения для дискриминантов: $\delta_1 = 12.81SH1 + 1.46SOE1 - 69.95$, $\delta_2 = 10.84SH1 + 1.99SOE1 - 78.15$. Для больного из популяции больных пневмонией с показателями $SH1 = 5.5$ и $SOE1 = 40$ получаем значения дискриминанта $\delta_1 = 59 < \delta_2 = 61.2$, откуда следует, что процедура классификации для этого больного работает правильно. Однако правильная классификация осуществляется не всегда. Из матрицы классификации

	Процент	W_1	W_2
W_1	95.45	63	3
W_2	92.54	5	62
Всего	93.98	68	65

узнаем, что трое больных бронхитом имеют настолько высокие показатели $SH1$ и $SOE1$, что процедура классификации относит их к больным пневмонией, а пять больных пневмонией по своим показателям более близки к больным бронхитом. Процент 95.45 носит специальное название чувствительности (процент правильной идентификации более легких больных), а 92.54 специфичности (процент правильной идентификации более тяжелых больных). Общий процент правильной классификации равен 93.98.

Пошаговый дискриминантный анализ

Задача пошагового дискриминантного анализа состоит в том, чтобы из множества признаков выделить такое подмножество, по которому можно построить наилучшее разделение популяций. Наилучшее разделение соответствует такому, при котором вероятность или стоимость ошибочной классификации была минимальной среди всех подмножеств.

Наиболее распространенный метод отбора переменных основан на так называемой Λ-статистике Уилкса.

$$\Lambda(\mathbf{X}) = \frac{W(\mathbf{X})}{T(\mathbf{X})},$$

где $\mathbf{X} = (X_1, \ldots, X_p)^T$ вектор p признаков, $W(\mathbf{X})$ – внутренняя обобщенная дисперсия, вычисляемая как определитель матрицы внутригрупповых перекрестных произведений, $T(\mathbf{X})$ – общая обобщенная дисперсия как опре-

делитель матрицы перекрестных произведений. Статистика Λ принимает значения от нуля до единицы: большие значения указывают на слабое разделение между группами, в то время как малые, наоборот, на хорошее разделение между группами, по крайней мере между некоторыми из них.

$$W(\mathbf{X}) = \{W_{st}\}_{s,t=1}^{p} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}^{s} - \bar{x}_i^{s})(x_{ij}^{t} - \bar{x}_i^{t}),$$

$$T(\mathbf{X}) = \{T_{st}\}_{s,t=1}^{p} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}^{s} - \bar{x}^{s})(x_{ij}^{t} - \bar{x}^{t}).$$

Мультипликативное приращение

$$\Lambda_u(X) = \frac{\Lambda(U,\mathbf{X})}{\Lambda(\mathbf{X})},$$

получаемое при добавлении переменной U к переменным $X_1, \ldots, X_p$, называется *частной Λ-статистикой.* Для проверки значимости изменений в $\Lambda(\mathbf{X})$ при добавлении переменной U используется статистика

$$F = \frac{n-k-p}{k-1} \cdot \frac{1-\Lambda_u(\mathbf{X})}{\Lambda_u(\mathbf{X})}$$

как средство отбора переменных, где k - число групп. При включении переменной U в множество $X_1, \ldots, X_p$ она называется статистикой F-включения, а при исключении переменной U из множества $U, X_1, \ldots, X_p$ статистикой F-исключения. Обозначим через $a = k-1$, $b = n-1-(p+k)/2$,

$$c = \begin{cases} \sqrt{\frac{p^2a^2-4}{p^1+a^2-5}} & \text{если } p^2+a^2 \neq 5, \\ 1, & \text{если } p^2+a^2 = 5 \end{cases}.$$

$$\text{Статистика} \quad F = \frac{df_1}{df_2} \cdot \frac{1-\Lambda^{1/c}}{\Lambda^{1/c}}$$

имеет приблизительно распределение Фишера с $df_1 = pa$ и $df_2 = bc + 1 - pa/2$ степенями свободы [Pao, 1957].

Чтобы выбрать переменную для первого шага дискриминантного анализа, рассмотрим F- значения, построенные для всех p компонент. Если все они меньше некоторого принятого минимального значения, которое называется F-включения, то средние значения внутри популяций считаются неотличимыми, и ни одна из переменных не вносит весомого вклада в разделение популяций. Иначе переходим к первому шагу.

Шаг 1. Из совокупности переменных, значение F-включения которых больше принятого минимума F-включения, выбирается переменная X_{j_1} с наибольшим значением F-включения. Она считается первой. Для каждой популяции W_i оценивается коэффициент и определяется постоянная дискриминантной функции. Кроме того, вычисляется таблица классификации, $F-$аппроксимация U- статистики. Для каждой из переменных X_j вычисляется значение F- включения с $k-1$ и $n-k-1$ степенями свободы. Таким образом проверяется гипотеза

$$H_0 : \mu^1_{j\cdot j_1} = \ldots = \mu^k_{j\cdot j_1},$$

где $\mu^i_{j\cdot j_1}$ среднее условного распределения X_j в популяции W_i при фиксированном значении переменной X_{j_1}. Если все F- включения меньше его минимума (ни одна из переменных не вносит значимого различия в популяции), то выполняется шаг S, в противном случае процедура переходит к выполнению второго шага.

Шаг 2. Выбирается переменная X_{j_2}, для которой значение статистики F-включения максимально, то есть та, по которой популяции отличаются наибольшим образом. Вычисляются оценки для двух коэффициентов и постоянные дискриминантных функций для каждой популяции W_i, таблица результатов классификации и F- аппроксимация U статистики.

Для переменных X_{j_1} и X_{j_2} ищутся значения статистик F -удаления с $k-1$ и $n-k-1$ степенями свободы. Таким образом проверяются гипотезы

$H_0 : \mu^1_{j_1 \cdot j_2} = \ldots = \mu^k_{j_1 \cdot j_2}$ и $H_0 : \mu^1_{j_2 \cdot j_1} = \ldots = \mu^k_{j_2 \cdot j_1}$. Во второй гипотезе значение статистики F больше F-включения, которая по умолчанию больше F-удаления, а первая может быть не отвергнута, в таком случае переменная X_{j_1} не вносит значимого различия в популяции при заданном значении переменной X_{j_1} и должна быть удалена из классифицирующих переменных.

Затем для каждой не включенной переменной вычисляется статистика F - включения с $k-1$ и $n-k-2$ степенями свободы. С ее помощью проверяется гипотеза $H_0 : \mu^1_{j \cdot j_1 j_2} = \ldots = \mu^k_{j \cdot j_1 j_2}$. Если все F включения меньше минимума, то выполняется шаг S, в противном случае шаг 3.

Шаг 3. А) Обозначим через L множество из l переменных, предварительно отобранных для построения процедуры классификации. Если значение F- удаления меньше принятого минимума, то переменная, для которой это значение минимально, удаляется из множества L. Выполняется Б) с заменой l на $l-1$. Если некоторые значения F для переменных, не входящих во множество L, больше принятого F-включения, то переменная с максимальным значением включается в L и l заменяется на $l+1$.

Б) Для каждой популяции W_i оцениваются дискриминантные функции, вычисляются результаты классификации и значения F аппроксимации U статистики. Затем для переменных из множества L вычисляются F-удаления $H_0 : \mu^1_{s \cdot (l-1)} = \ldots = \mu^k_{s \cdot (l-1)}$ для каждой переменной X_s из L при фиксированных остальных $(l-1)$ элементов. Затем вычисляется значение F -включения для переменных, не включенных в L. $H_0 : \mu^1_{j \cdot (l)} = \ldots = \mu^k_{j \cdot (l)}$. $\mu^i_{i \cdot (l)}$ есть условное математическое ожидание переменной X_j в W_i при фиксированных значениях переменных из L. $X_j \notin L$.

Процедура переходит к финальному шагу, когда статистики F-включения меньше минимума или когда все переменные оказываются включенными, а значения F - удаления меньше минимума, то выполняется шаг

S.

Шаг S. Вычисляются апостериорные вероятности. На их основании производится классификация, строится таблица классификации.

ПРИМЕР (продолжение). Рассмотрим процедуру пошагового дискриминантного анализа для решения задачи выделения наиболее информативных признаков и диагностики бронхита и пневмонии на основе четырех показателей: X_1 – возраст, X_2 – серосодержащие белки в востановительной форме, X_3 – лейкоциты, X_4 – скорость оседания эритроцитов. По этим признакам анализируется популяция из 133 больных, среди которых 66 бронхитом W_1 и 67 больных пневмонией W_2. Построим дискриминантную процедуру, которая бы позволила по этим четырем признакам диагностировать заболевание. Определим минимум F-включения равным 4.0, а минимум F-удаления равным 3.9.

Шаг 0. Для всех четырех переменных вычисляем значения F-включения с 1 и 131 степенями свободы.

	X_1	X_2	X_3	X_4
F	0.25	86.75	41.70	238

По переменной X_4 (СОЭ) строим дискриминанты: $d_1 = 1.37x_4 - 23.15$ и $d_2 = 1.91x_1 - 44.65$. Если дискриминант d_1 больше d_2, то пневмония, иначе бронхит. На основании этого получены следующие результаты классификации.

	W_1	W_2
W_1	56	10
W_2	4	63

Значение F-удаления совпадает на этой стадии со значением F-включения и равно 238. Следовательно, переменная X_4 не удаляется из множества участвующих в дискриминации признаков. Для переменных X_1, X_2 и X_3 вычисляются значения F-включения с 1 и 130 степенями свободы, то есть проверяется, не будут ли равны средние значения этих признаков в обеих популяциях в случае, когда устранено влияние переменной X_4.

	X_1	X_2	X_3
F	0.08	25.9	1.86

Одно значение больше установленного минимума, поэтому выбираем в качестве следующей переменной X_2 - серосодержащие восстановительные белки.

Шаг 2. Строим дискриминанты по этим двум переменным. $d_1 = 1.46X_4 + 12.8X_2 4 - 69.9$ и $d_2 = 1.99x_4 + 10.8X_2 - 78.1$. Результаты классификации выглядят следующим образом:

	W_1	W_2
W_1	63	3
W_2	5	62

Для переменных X_2 и X_4 вычисляем F-удаления. Соответственно они равны 25.9 и 134, оба значения больше 3.9, следовательно, ни одна из этих переменных не удаляется. Для переменных X_1 и X_3, не участвующих в классификации, значения F-включения равны соответственно 56.8 и 4.48. Следующая переменная X_1 - возраст.

Шаг 3. $d_1 = 1.3X_4 + 53.2X_2 + 2.6X_1 - 269.5$ и $d_2 = 1.8X_4 + 46.3X_2 + 2.38X_1 - 232.2$. Результаты классификации:

	W_1	W_2
W_1	63	3
W_2	3	64

Значения F -удаления равны 79.07, 93.74, 56.8, все больше 3.9. Значение F-включения для переменной X_3 равно 4.6, оно больше 3.9.

Шаг 4. $d_1 = 1.35X_4 + 53.55X_2 + 2.6X_1 - 0.32X_3 - 270$ и $d_2 = 1.85x_4 + 46.4X_2 + 2.29x_1 - 0.05X_3 - 232$. Значения F-удаления:

	X_4	X_2	X_1	X_3
F	49	98	57	4.6

Ни одна из переменных не удаляется, поэтому окончательные результаты классификации выглядят следующим образом:

	W_1	W_2
W_1	65	1
W_2	3	64

Значение F-аппроксимации статистики Уилкса $F = (4, 128) = 129.3575, p < 0.0000$, то есть переменные вносят значимый вклад в дискриминацию двух популяций, то есть по данным признакам можно диагностировать бронхит и пневмонию.

Для того чтобы сделать обработку по дискриминантному анализу в $StatSoft$, нужно запустить программу *Statistics - Multivariate Exploratory Techniques - Discriminant Analysis*. Далее нужно выбрать переменные ($Variables$). При необходимости привлечения пошаговой процедуры ставится галочка в *Advanced options* (*stepwise analysis*). Далее OK. В таком случае нужно будет выбрать метод: $Forward$ или *Backward* в зависимости от того, происходит ли отбор информативных или исключение малоинформативных признаков. Дискриминантные коэффициенты можно получить из $Classification\ functions$, матрицу классификаций из $Classification\ matrix$, апостериорные вероятности соответственно из *Posterior probabilities*. Переход от байесовской процедуры к эвристической возможен за счет выбора априорных вероятностей *A priori classification probabilities*.

7.3. Классификация в случае биномиальных распределений

Пусть имеются p дихотомических признаков $X_1, \ldots, X_p$, значения которых состоят только из нулей и единиц:

$$X_j = \begin{cases} 1, & "yes \\ 0, & "no" \end{cases}, \quad j = 1, \ldots, p,$$

например, ответы "да"и "нет"на ряд вопросов или наличие и отсутствие некоторых симптомов. Решается задача отнесения объекта к одной из k популяций $W_1, \ldots, W_k$ с биномиальными распределениями с известными априорными вероятностями $q_1, \ldots, q_k$. Положим

$$P(X_j = 1|W_i) = p_{ij} \quad \text{и} \quad P(X_j = 0|W_i) = 1 - p_{ij},$$

$i = 1, \ldots,\ k, j = 1, \ldots, p$. Тогда закон распределения X_j для популяции W_i имеет вид

$$f_i(x_j) = p_{ij}^{x_j}(1 - p_{ij})^{1-x_j}, \quad x_j \in \{0, 1\}.$$

При независимости признаков $X_1, \ldots, X_p$ совместный закон распределения

имеет вид

$$f_i(\mathbf{x}) = f_i(x_1) \cdot \ldots \cdot f_i(x_p), \quad i = 1, \ldots, k, \quad \mathbf{x} = (x_1, \ldots, x_p)^T.$$

Вектор $\mathbf{x}$ относится к популяции W_i с максимальной величиной апостериорной вероятности, которая при одинаковой стоимости ошибочной классификации имеет вид:

$$P(W_i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{\sum\limits_{m=1}^{k} q_m f_m(\mathbf{x})}.$$

Пусть n_i число индивидов из популяции W_i, из них n_{ij} имеют j-ый симптом. Тогда оценка $\hat{p_{ij}} = n_{ij}/n_i$. Если априорные вероятности неизвестны, то их оценками являются $\hat{q}_i = n_i/n$, где $n_1 + \ldots + n_k = n$.

Поскольку в большинстве случаев предположение о независимости не выполняется, то когда все априорные вероятности равны, для классификации используется процедура, которая для любой комбинации симптомов вычисляет соответствующую долю объектов в каждой выборке. Новый объект относится к популяции, выборка из которой имеет наибольшую долю объектов с такой комбинацией. Например, $k = 2, p = 3$.

Комбинации	000	001	010	011	100	101	110	111
W_1	0.10	0.10	0.20	0.10	0.20	0.10	0.10	0.10
W_2	0.20	0.15	0.10	0.15	0.10	0.20	0.05	0.05

Поступающий объект с комбинацией симптомов 011 будет отнесен к популяции W_2, так как $0.15 > 0.10$. Неудобство этого метода заключается в том, что для каждой популяции надо вычислять 2^p параметров.

ПРИМЕР 1. Рассмотрим популяции (табл.3) W_1 (слова, означающие число "1 на $n_1 = 9$ языках) и W_2 (слова, означающие число "3 на тех же $n_2 = 9$ языках). В качестве дихотомических признаков X_j, $j = 1, 2, \ldots, 12$, будем использовать факт наличия в слове некоторых однотипных звуков. Имеем $k = 2$ число популяций, $p = 12$ число признаков. Пусть n_{ij} – число слов из популяции W_i, содержащих звук X_j, оценка p_{ij} имеет вид:

$$\widehat{p_{ij}} = n_{ij}/n_i, \quad i = 1, 2; \quad j = 1, 2, \ldots, 12.$$

Выражение $n_{14} = 3$ означает, что в популяции слов, означающих число "один три слова содержат буквы "к,х,г"(вахид, йек, нынг).

Оценками априорных вероятностей являются $\widehat{q}_i = n_i/n, \quad i = 1, 2, \ldots, k$, где $n = \sum_{l=1}^{k} n_l$ (табл.4). В нашем случае априорные вероятности равны $\widehat{q}_1 = 9/18 = 0.5; \quad \widehat{q}_2 = 9/18 = 0.5$.

Например, слову "вахид"соответствует вектор $X = (101110010000)$. Вероятность его появления при условии W_1 равна

$$\widehat{f}_1(X) = \frac{2}{9} \cdot \frac{8}{9} \cdot \frac{1}{9} \cdot \frac{3}{9} \cdot \frac{7}{9} \cdot \frac{6}{9} \cdot \frac{8}{9} \cdot \frac{8}{9} \cdot \frac{6}{9} \cdot \frac{9}{9} \cdot \frac{8}{9} \cdot \frac{9}{9} = 0.002.$$

язык	один	три
арабский	вахид	саляса
персидский	йек	сэ
турецкий	бит	юч
японский	ити	сан
амхарский	анд	сост
тайский	нынг	сам
грузинский	эрти	сами
китайский	и	сань
аккадский	иштен	шалаш

Таблица 3. Названия цифр 1 и 3 в восточных языках.

Если компонента $X_j = 1$, то соответствующий сомножитель равен $\widehat{p}_{1j}$, если $X_j = 0$, то сомножитель равен $1 - \widehat{p}_{1j}$. Аналогично вероятность появления вектора $X = (101110010000)$ при условии W_2 равна

$$\widehat{f}_2(X) = \frac{6}{9} \cdot \frac{9}{9} \cdot \frac{0}{9} \cdot \frac{1}{9} \cdot \frac{1}{9} \cdot \frac{8}{9} \cdot \frac{0}{9} \cdot \frac{1}{9} \cdot \frac{3}{9} \cdot \frac{8}{9} \cdot \frac{9}{9} \cdot \frac{8}{9} = 0.$$

Вероятность получения вектора $X = (101110010000)$ равна

$$P(X) = \widehat{q}_1 \cdot \widehat{f}_1(X) + \widehat{q}_2 \widehat{f}_2(X) = 0.5 \cdot 0.002 + 0.5 \cdot 0 = 0.001.$$

Апостериорные вероятности равны

$$P(W_1|X) = \frac{\widehat{q}_1 \cdot \widehat{f}_1(X)}{P(X)} = \frac{0.5 \cdot 0.002}{0.001} = 1, \quad P(W_2|X) = \frac{\widehat{q}_2 \cdot \widehat{f}_2(X)}{P(X)} = \frac{0}{0.001} = 0.$$

Следовательно, слово "вахид"правильно классифицируется как слово из популяции W_1 единиц.

Результаты 100%-правильной классификации остальных слов сводятся в таблицу:

	к популяции W_1	к популяции W_2
из популяции W_1	9	0
из популяции W_2	0	9

j	1	2	3	4	5	6	7	8	9	10	11	12
X_j	а я	б п	в ф	г к х	д т	е э	с з ш ч	й ы и	л м н	о	р	у ю
вахид	1	0	1	1	1	0	0	1	0	0	0	0
йек	0	0	0	1	0	1	0	1	0	0	0	0
бит	0	1	0	0	1	0	0	1	0	0	0	0
ити	0	0	0	0	1	0	0	1	0	0	0	0
анд	1	0	0	0	1	0	0	0	1	0	0	0
нынг	0	0	0	1	0	0	0	1	1	0	0	0
эрти	0	0	0	0	1	1	0	1	0	0	1	0
ити	0	0	0	0	1	0	0	1	0	0	0	0
иштен	0	0	0	0	1	1	1	1	1	0	0	0
саляса	1	0	0	0	0	0	1	0	1	0	0	0
сэ	0	0	0	0	0	1	1	0	0	0	0	0
юч	0	0	0	1	0	0	1	0	0	0	0	1
сан	1	0	0	0	0	0	1	0	1	0	0	0
сост	0	0	0	0	1	0	1	0	0	1	0	0
сам	1	0	0	0	0	0	1	0	1	0	0	0
сами	1	0	0	0	0	0	1	1	1	0	0	0
сань	1	0	0	0	0	0	1	0	1	0	0	0
шалаш	1	0	0	0	0	0	1	0	1	0	0	0
n_{1j}	2	1	1	3	7	3	1	8	3	0	1	0
n_{2j}	6	0	0	1	1	1	9	1	6	1	0	1

Таблица 4. Вычисление частот появления звуков в слове.

ПРИМЕР 2. Рассмотрим, например, две популяции.

W_1	W_2
$n_1 = 5$	$n_2 = 13$
Россия	Италия
Болгария	Канада
Венгрия	Австралия
Румыния	Австрия
Польша	Бельгия
	Дания
	Мексика
	Нидерланды
	США
	Финляндия
	Франция
	Швеция
	Япония

Четыре ($p = 4$) дихотомических признака

- X_1: 0 – низкая (< 300) или 1 – высокая смертность от заболеваний системы кровообращения;
- X_2: 0 – низкая (< 189) или 1 – высокая смертность от злокачественных новообразований;
- X_3: 0 – низкая (< 10) или 1 – высокая смертность от инфекционных заболеваний;
- X_4: 0 – низкая (< 50) или 1 – высокая смертность от несчастных случаев и других внешних воздействий.

n_{ij} – количество стран из популяции W_i с высокими показателями смертности X_j. $\hat{p}_{ij} = \frac{n_{ij}}{n_i}$ – оценка вероятности высокого показателя смертности X_j в популяции W_i.

X_j	1	2	3	4
n_{1j}	5	3	2	5
n_{2j}	5	9	3	7

X_j	1	2	3	4
$\hat{p}_{1j}$	1	0.6	0.4	1
$\hat{p}_{2j}$	0.385	0.692	0.230	0.538

Выясним, к какой популяции W_1 или W_2 будет отнесена Германия со значениями признаков $X = (1100)^T$ результате процедуры классификации. В качестве оценок априор-

ных вероятностей рассмотрим $\hat{q}_1 = n_1/n = 13/18 = 0.722$ и $\hat{q}_2 = n_2/n = 5/18 = 0.278$. Вероятность получить значение вектора $X = (1100)^T$ при условии W_1 равна

$$\hat{f}_1(1,1,0,0) = \hat{p}_{11}\hat{p}_{12}(1-\hat{p}_{13})(1-\hat{p}_{14}) = 1 \cdot 0.6 \cdot (1-0.4) \cdot (1-1) = 0.$$

Вероятность получить значение вектора $X = (1100)^T$ при условии W_2 равна

$$\hat{f}_2(1,1,0,0) = \hat{p}_{21}\hat{p}_{22}(1-\hat{p}_{23})(1-\hat{p}_{24}) = 0.385 \cdot 0.692 \cdot (1-0.230) \cdot (1-0.538) = 0.0945.$$

Оценки апостериорных вероятностей имеют вид:

$$P(W_1|1,1,0,0) = \frac{\hat{q}_1\hat{f}_1(1,1,0,0)}{\hat{q}_1\hat{f}_1(1,1,0,0) + \hat{q}_2\hat{f}_2(1,1,0,0)} = \frac{0}{0.0683+0} = 0,$$
$$P(W_2|1,1,0,0) = \frac{\hat{q}_2\hat{f}_2(1,1,0,0)}{\hat{q}_1\hat{f}_1(1,1,0,0) + \hat{q}_2\hat{f}_2(1,1,0,0)} = \frac{0.0683}{0.0683+0} = 1.$$

Наибольшая апостериорная вероятность достигается в случае популяции W_2, т.е. процедура классификации по показателям смертности характеризует Германию, в свое время объединившую свои восточную и западную части, как ее западную часть.

8. Факторный анализ

8.1. Линейные преобразования, приводящие к некоррелированным компонентам

Рассмотрим два центрированных и нормированных признака X и Y с коэффициентом корреляции ϱ. На рис.8 представлена двумерная диаграмма признаков: X - объемы продаваемой водки, Y - продолжительность жизни с коэффициентом корреляции $\widehat{\varrho} = r = -0.934$. В теме о множественной регрессии уже использовалось свойство некоррелированности остатка с вычитаемой переменной.

Если перейти к новым координатам в результате линейного преобразования:

$$\begin{cases} X_0 = 1 \cdot X + 0 \cdot Y \\ Y_0 = -r \cdot X + 1 \cdot Y \end{cases}, \tag{84}$$

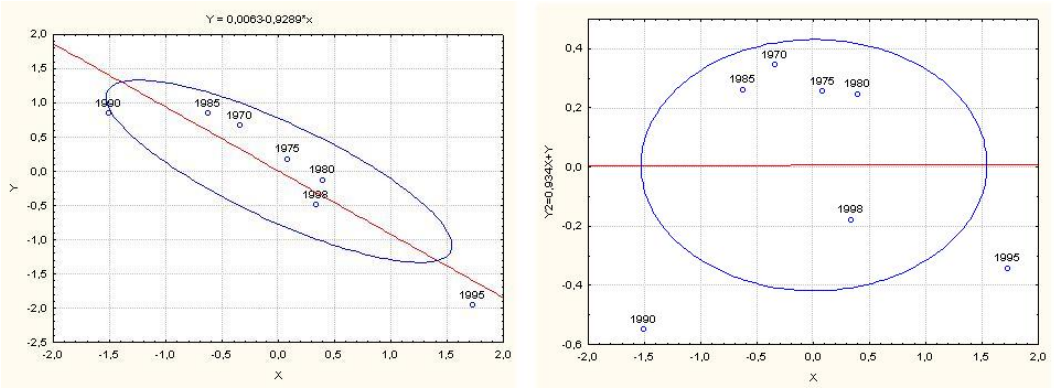


Рис. 8. Двумерные диаграммы между объемом продажи водки X и продолжительностью жизни Y (данные нормированы и центрированы, $r = -0.934$), а также между X и линейным преобразованием $Y - rX$.

то мы получим две некоррелированные компоненты X_0 и Y_0 (рис.8).

Линейных преобразований, приводящих к некоррелированным компонентам, может быть сколько угодно. Но среди них можно выделить обладающее рядом замечательных свойств преобразование вида:

$$\begin{cases} X_1 = \frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y \\ Y_1 = \frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y \end{cases}. \tag{85}$$

Новые случайные величины X_1 и Y_1 называются *главными компонентами.* В случае центрированных и нормированных признаков главные компоненты имеют нулевые средние, а их дисперсии равны соответственно

$$DX_1 = D\left(\frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y\right) = \frac{DX}{2} + \frac{DY}{2} + EXY = 1 + \varrho,$$
$$DY_1 = D\left(\frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y\right) = \frac{DX}{2} + \frac{DY}{2} - EXY = 1 - \varrho.$$

При отрицательном коэффициенте корреляции система (85) чуть-чуть видоизменяется, что однако не влияет на свойства системы.

$$\begin{cases} X_1 = -\frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y \\ Y_1 = \frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y \end{cases}. \tag{86}$$

Свойства преобразования (85):

1. главные компоненты X_1 и Y_1 некоррелированы, так как

$$EX_1Y_1 = E\left(\frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y\right)\left(\frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y\right) = \\ = \frac{1}{2}\left(EX^2 - EY^2\right) = 0.$$

2. преобразование (85) сохраняет суммарную дисперсию, то есть

$$DX_1 + DY_1 = D\left(\frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y\right) + D\left(\frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y\right) = \\ = \frac{1}{2}\left(DX + DY + 2EXY\right) + \\ + \frac{1}{2}\left(DX + DY - 2EXY\right) = D(X) + D(Y).$$

Заметим, что для преобразования (84) это свойство не выполняется (проверить самостоятельно).

3. преобразование (85) сохраняет расстояния между точками:

$$X_1^2 + Y_1^2 = \left(\frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y\right)^2 + \left(\frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y\right)^2 = \\ = \left(\frac{X^2}{2} + \frac{Y^2}{2} + XY\right) + \left(\frac{X^2}{2} + \frac{Y^2}{2} - XY\right) = X^2 + Y^2.$$

Таким образом в системе (86) мы получили представление наших данных совсем в других координатах. Первая координата, которая увеличивается при увеличении продолжительности жизни и уменьшается при увеличении объемов продажи водки, может быть интерпретирована как фактор здоровья населения за счет ограничения продажи алкоголя. В этом смысле самыми благополучными являются годы 1990 и 1985, а самым неблагополучным 1995. Вторая координата, которая увеличивается одновременно с увеличением и продолжительности жизни, и продажами алкоголя, может быть интерпретироваться как фактор позитивного влияния алкоголя на продолжительность жизни. Это весьма сомнительно, поэтому рассмотрим,

насколько значимым является *вклад второй компоненты в общую дисперсию.*

Поскольку данные нормированы, то дисперсии равны $D(X) = D(Y) = 1$. Общая дисперсия равна 2. При отрицательном коэффициенте корреляции дисперсия главной компоненты $D(X_1) = 1 - \varrho$, следовательно, ее вклад в общую дисперсию равен

$$\frac{D(X_1)}{D(X_1) + D(Y_1)} = \frac{1 - \varrho}{2} = \frac{1 - (-0.934)}{2} = 0.967.$$

Поэтому оставшейся второй компоненте остается 0.033. Вклад 3.3% не настолько существенен, чтобы его считать значимым.

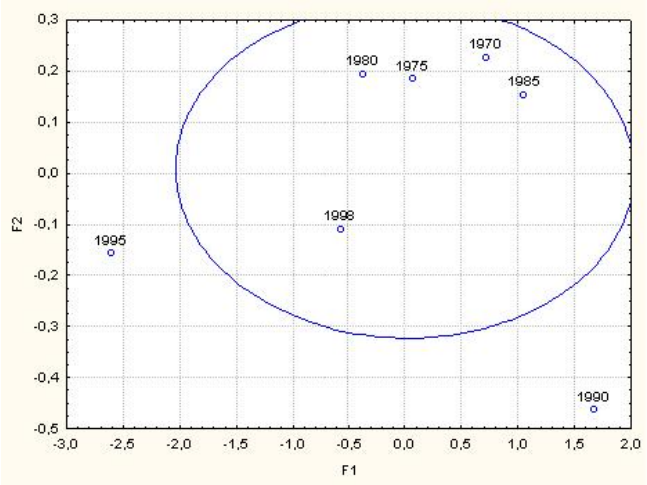


Рис. 9. Двумерная диаграмма некоррелированных главных компонент.

8.2. Метод главных компонент в случае $k \geq 2$ переменных

Рассмотрим ситуацию с числом признаков, большим двух. В большинстве случаев эти признаки коррелированы, поскольку являются результатами частных измерений некоторых интегральных характеристик. В медицине эти скрытые признаки чаще всего отражают степень тяжести того или иного заболевания и мощность защитных сил организма и складываются из влияния целого комплекса характеристик. Часто оказывается, что эти

интегральные характристики описываются главными компонентами. Идея метода главных компонент состоит в том, чтобы заменить k-мерную случайную величину при наименьшей потере информативности на m- мерную $(m < k)$.

Что означает наименьшая потеря информации? Эффективность любого выбора зависит от того, в какой степени эти m линейных функций дают возможность реконструировать или восстановить k первоначальных величин. Один из методов реконструкции этой первоначальной случайной величины состоит в построении ее наилучшего предиктора на основе m линейных функций. Оказывается, что наилучшим выбором линейных функций является выбор первых m главных компонент.

Определим понятие главной компоненты. Пусть имеется k признаков $X_1, \ldots, X_k$. Первой главной компонентой Y_1 называется сохраняющая расстояние между точками линейная комбинация исходных признаков

$$Y_1 = \alpha_{11}X_1 + \ldots + \alpha_{k1}X_k,$$

где коэффициенты $\alpha_{11}, \ldots, \alpha_{k1}$ выбираются таким образом, чтобы дисперсия $D(Y_1) = \lambda_1$ была максимальной. Это соответствует тому, что по первой главной компоненте индивиды должны отличаться наибольшим образом. Вторая главная компонента также является линейной комбинацией исходных признаков:

$$Y_2 = \alpha_{12}X_1 + \ldots + \alpha_{k2}X_k,$$

где коэффициенты $\alpha_{12}, \ldots, \alpha_{k2}$ выбираются таким образом, что компоненты Y_1 и Y_2 некоррелированы, а дисперсия $D(Y_2) = \lambda_2$ является максимальной из всех линейных комбинаций, некоррелированных с Y_1, то есть вторая компонента должна нести наибольшую новую информацию, не имеющую отношения к первой главной компоненте. Аналогично строятся остальные

главные компоненты

$$Y_j = \sum_{i=1}^{k} \alpha_{ij} X_i,\ j = 1, \ldots, k.$$

Суммарная дисперсия остается неизменной:

$$V = D(X_1) + \ldots D(X_k) = \lambda_1 + \ldots + \lambda_k.$$

Значимость главных компонент определяется долей объясняемой ими дисперсии, равной $\frac{\lambda_i}{V} \cdot 100\%$.

Факторами называются нормированные главные компоненты: $Y_j/\sqrt{\lambda_i}$.

Для того чтобы сделать факторную обработку в $StatSoft$, нужно запустить программу *Statistics - Multivariate Exploratory Techniques - Factor Analysis*. Далее нужно выбрать переменные, по которым строятся факторы (*Variables*). Нажимаем OK.

В окошке $Max.no.of factors$ нужно указать необходимое для анализа число факторов, оно должно быть больше нуля и не больше количества участвующих в анализе признаков.

В окошке *Mini. eigenvalue* лучше сразу поставить 0. Читаем, сколько переменных и сколько наблюдений участвует в анализе. Далее OK. Для правильного чтения результатов необходимыми являются сведения, излагаемые далее.

8.3. Вычисление коэффициентов главных компонент

Пусть признаки $X = (X_1, \ldots, X_k)^T$ центрированы $EX_i = 0$ и имеют ковариационную матрицу $\Sigma = EXX^T$. Обозначим через A_i собственные векторы матрицы Σ, соответствующие собственным числам λ_i

$$\Sigma A_i = \lambda_i A_i, \quad A_i^T A_j = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \quad i = 1, \ldots, k. \tag{87}$$

Упорядочим собственные числа по убыванию: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$. Соберем все собственные вектора в одну ортогональную матрицу

$$\mathcal{A} = [A_1, \ldots, A_k] = \begin{bmatrix} a_{11}, & \ldots, & a_{1k} \\ . & . & . \\ a_{k1}, & \ldots, & a_{kk} \end{bmatrix}, \quad \mathcal{A}^T\mathcal{A} = I = \begin{bmatrix} 1, & \ldots, & 0 \\ . & . & . \\ 0, & \ldots, & 1 \end{bmatrix}$$

и определим главную компоненту как

$$Y_j = A_j^T X = \sum_{i=1}^{k} a_{ij} X_i. \tag{88}$$

Для всех главных компонент справедливо выражение:

$$Y = \begin{bmatrix} Y_1 \\ . \\ Y_k \end{bmatrix} = \begin{bmatrix} A_1^T X \\ . \\ A_k^T X \end{bmatrix} = \mathcal{A}^T X, \quad \text{откуда} \quad X = \mathcal{A} Y. \tag{89}$$

Теорема (о некоррелированности главных компонент и инвариантности суммарной дисперсии).

$$1.\ EY_i Y_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \qquad 2.\ \sum_{i=1}^{k} DX_i = \sum_{j=1}^{k} DY_j.$$

Доказательство.

$$1. EY_i Y_j = EY_i Y_j^T \overset{(88)}{=} EA_i^T X (A_j^T X)^T = A_i^T \Sigma A_j \overset{(87)}{=}$$
$$= A_i^T \lambda_j A_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j. \end{cases}$$

Следовательно, дисперсия главной компоненты равна соответственному собственному числу ковариационной матрицы: $DY_j = \lambda_j$. Из упорядоченности собственных чисел получаем упорядоченные по убыванию дисперсии главные компоненты: $DY_1 \geq DY_2 \geq \ldots \geq DY_k$.

$$2.\ DX_1 + \ldots + DX_k = \sum_{i=1}^{k} EX_i^2 = EX^T X = \overset{(89)}{=}$$
$$= EY^T \mathcal{A}^T \mathcal{A} Y = EY^T Y = \sum_{j=1}^{k} EY_j^2 = DY_1 + \ldots + DY_k.$$

8.4. Факторные нагрузки и формула восстановления

Корреляция $\beta_{ij} = cor(X_i, Y_j)$ между признаком X_i и главной компонентой Y_j называется *факторной нагрузкой*. Пусть σ_i – стандартное отклонение X_i, λ_j – дисперсия главной компоненты Y_j. Тогда факторная нагрузка имеет вид:

$$\beta_{ij} = cor(X_i, Y_j) = \frac{EX_iY_j}{\sigma_i\sqrt{\lambda_j}} \overset{(89)}{=} \frac{E\sum\limits_{t=1}^{k}\alpha_{it}Y_tY_j}{\sigma_i\sqrt{\lambda_j}} = \frac{\alpha_{ij}\sqrt{\lambda_j}}{\sigma_i} \qquad (90)$$

Формулу восстановления X_i по главным компонентам можно выразить через факторы $G_j = Y_j/\sqrt{\lambda_j}$:

$$X_i = \sum_{j=1}^{k}\alpha_{ij}Y_j = \sigma_i\sum_{j=1}^{k}\frac{\beta_{ij}}{\sqrt{\lambda_j}}G_j\sqrt{\lambda_j} = \sigma_i\sum_{j=1}^{k}\beta_{ij}G_j\,.$$

При восстановлении переменных по m главным компонентам, меньшему количества исходных признаков k, значения признаков могут восстанавливаться с ошибками. Чем больше вклад используемых в восстановлении главных компонент, тем меньше ошибки восстановления.

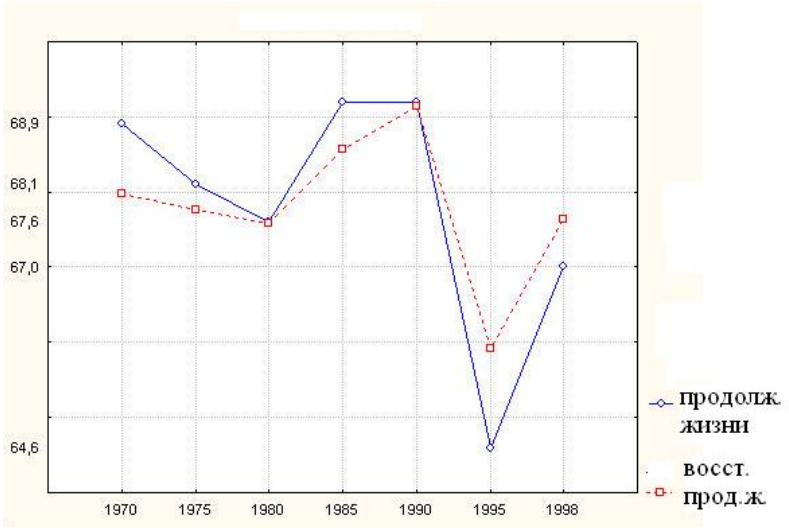


Рис. 10. Восстановление средней продолжительности жизни по первому и второму факторам. $\widehat{L} = (0.896f_1 + (-0.398f_2)) \cdot 1.64 + 67.8$,

факторные нагрузки

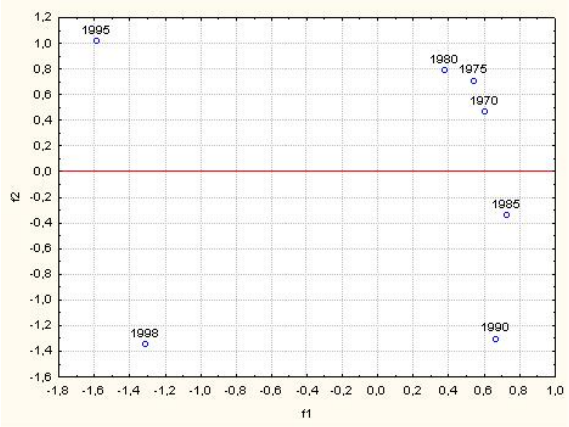
	Factor 1	Factor 2
L	0.896	−0.398
M	−0.815	−0.564
P	0.905	−0.045
A	−0.847	−0.486
V	−0.772	0.613
Дисп.гл.комп. λ	3.60	1.06
вклад в дисп.	72%	21.8%

Значения факторов

годы	f_1	f_2
1970	0.600	0.465
1975	0.540	0.710
1980	0.375	0.790
1985	0.724	−0.339
1990	0.662	−1.301
1995	−1.587	1.019
1998	−1.314	−1.345

ПРИМЕР 1. Признаки: L – средняя продолжительность жизни; M – количество чиновников; A – количество автомобилей; P – доходы бедных; V – объемы продажи водки.

Для того чтобы увидеть матрицу факторных нагрузок, нужно открыть вкладку *Loading* и нажать кнопку $Summary : Factor\ loadings$. Для получения значений факторов G_j нужно во вкладке *Scores* выбрать $Factor\ Scores$ или $Save\ factorscores$.



Вклад первого фактора равен 72%. Значения в i-й строке и j-м столбце соответствуют коэффициенту корреляции между i-м признаком и j-й главной компонентой. Чем больше первый фактор, тем больше продолжительность жизни и доходы бедных, меньше чиновников и автомобилей и не много водки – фактор благополучия. С первым фактором имеет значимую корреляцию количество заключаемых браков.

Второй фактор менее значим, его вклад равен 22%. Большие значения второго фактора соответствуют тому, что много пьют, а автомобилей мало, соответственно

малые наоборот: автомобилей много, пьют мало. Этот фактор имеет значимо положительную корреляцию с числом разводов.

9. Канонические корреляции

9.1. Постановка задачи

Обычные коэффициенты корреляции используются для выявления линейной зависимости между двумя признаками X и Y. Если нужно выявить зависимость между признаком X_0 и набором признаков $X_1, \ldots, X_p$, то в качестве характеристики этой зависимости рассматривается множественный коэффициент корреляции, равный коэффициенту корреляции между X_0 и его наилучшим линейным предсказанием $\widehat{X_0} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

Этот принцип можно распространить на выявление зависимости между двумя наборами признаков $X_1, \ldots, X_r$ и $X_{r+1}, \ldots, X_{r+s}$. Математическая задача здесь состоит в том, чтобы найти такие нормированные линейные комбинации

$$U_1 = \beta_{10} + \beta_{11} X_1 + \ldots + \beta_{1r} X_r \quad \text{и} \quad V_1 = \alpha_{10} + \alpha_{11} X_{r+1} + \ldots + \alpha_{1s} X_{r+s},$$

чтобы корреляция $R_1 = cor(U_1, V_1)$ между U_1 и V_1 была максимальной. Случайные величины U_1 и V_1, а также их коэффициенты называются *каноническими.*

9.2. Метод нахождения канонических коэффициентов

Не умаляя общности, будем считать центрированными компоненты случайного вектора

$$X = (X_1, \ldots, X_r, X_{r+1}, \ldots, X_{r+s})^T$$

$EX_i = 0,\ i = 1, 2, \ldots, r+s,\ U_1 = (X_1, \ldots, X_r)^T, U_2 = (X_{r+1}, \ldots, X_{r+s})^T$. Обозначим через $\Sigma_{11} = EU_1U_1^T$, $\Sigma_{22} = EU_2U_2^T$ и $\Sigma_{12} = EU_1U_2^T$, $\Sigma_{21} = EU_2U_1^T$, $\Sigma_{12} = \Sigma_{21}$ ковариационные матрицы подвекторов, образующие общую ковариационную матрицу

$$\Sigma = EXX^T = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Канонические случайные величины обозначим через $V_1 = L^TU_1$, $V_2 = M^TU_2$, где $L = (l_1, \ldots, l_r)^T$, $M = (m_1, \ldots, m_s)^T$ векторы канонических коэффициентов. Поскольку

$$DV_1 = EV_1^2 = EL^TU_1U_1^TL = L^T\Sigma_{11}L = 1,$$
$$DV_2 = EV_2^2 = EM^TU_2U_2^TM = M^T\Sigma_{22}M = 1,$$
$$cov(V_1, V_2) = EV_1V_2 = EL^TU_1U_2^TM = L^T\Sigma_{12}M = M^T\Sigma_{21}L,$$

то экстремальная задача нахождения L и M таких, что

$$cor(V_1, V_2) = cov(V_1, V_2) \ \to \ max \quad \text{при} \quad DV_1 = DV_2 = 1 \tag{91}$$

сводится к задаче:

$$L^T\Sigma_{12}M \ \to \ max \quad \text{при} \quad L^T\Sigma_{11}L = 1 \text{ и } M^T\Sigma_{22}M = 1. \tag{92}$$

Для решения этой задачи используем метод множителей Лагранжа. Введем функцию Лагранжа

$$\mathcal{L}(L, M) = L^T\Sigma_{12}M - \frac{\lambda_1}{2}\left(L^T\Sigma_{11}L - 1\right) - \frac{\lambda_2}{2}\left(M^T\Sigma_{22}M - 1\right)$$

и построим систему нормальных уравнений:

$$\begin{cases} \Sigma_{12}M - \lambda_1\Sigma_{11}L = 0 \\ \Sigma_{21}L - \lambda_2\Sigma_{22}M = 0 \\ L^T\Sigma_{11}L = 1 \\ M^T\Sigma_{22}M = 1 \end{cases}$$

Умножая первое уравнение на L^T, а второе на M^T, получаем, что

$$\begin{cases} L^T\Sigma_{12}M - \lambda_1 L^T\Sigma_{11}L = 0 \\ M^T\Sigma_{21}L - \lambda_2 M^T\Sigma_{22}M = 0 \\ L^T\Sigma_{11}L = 1 \\ M^T\Sigma_{22}M = 1 \end{cases} \iff \quad \lambda_1 = \lambda_2 = cor(V_1, V_2) = \varrho\,.$$

Итак, имеем систему уравнений:

$$\begin{cases} \Sigma_{12}M - \varrho\Sigma_{11}L = 0 \\ \Sigma_{21}L - \varrho\Sigma_{22}M = 0 \end{cases}. \tag{93}$$

Умножим первое уравнение на $\Sigma_{21}\Sigma_{11}^{-1}$, а второе на ϱ, и сложим уравнения:

$$\begin{cases} \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}M - \varrho\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11}L = 0 \\ \varrho\Sigma_{21}L - \varrho^2\Sigma_{22}M = 0 \\ \quad \Longrightarrow \quad (\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \varrho^2\Sigma_{22})M = 0, \end{cases}$$

откуда получаем, что ϱ^2 является собственным корнем, а M собственным вектором уравнения

$$|\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \varrho^2\Sigma_{22}| = 0.$$

Аналогично в случае другого множества величин

$$(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \varrho^2\Sigma_{11})L = 0,$$

и ϱ^2 является собственным корнем, а L собственным вектором уравнения

$$|\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \varrho^2\Sigma_{11}| = 0.$$

Количество ненулевых корней равно рангу матрицы Σ_{12} В первом случае количество корней равно рангу матрицы Σ_{22}, то есть s, во втором рангу матрицы Σ_{11}, то есть r. Количество ненулевых корней равно m рангу

матрцы Σ_{12}, а кратности нулевых корней различны. Таким образом, имеем $m \leq min(r, s)$ канонических случайных величин $V_{1i} = L_i^T U_2$ и $V_{2i} = M_i^T U_2$, упорядоченных в соответствии с корнями $\varrho_1^2 \geq \varrho_2^2 \geq \ldots \geq \varrho_m^2$. Далее для удобства будем считать канонические корреляции ϱ_i положительными, иначе можно поменять знаки канонических коэффициентов M или L.

9.3. Свойства канонических случайных величин

$$1)\ cov(V_{1i}, V_{1j}) = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \quad cov(V_{2i}, V_{2j}) = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}.$$

$$2)\ \ cov(V_{1i}, V_{2j}) = \begin{cases} \varrho_i, & \text{при } i = j \leq m, \\ 0, & \text{при } i \neq j \text{ или } i = j > m \end{cases}$$

Обозначим через A симметричную матрицу $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, M_i и M_j собственные векторы, соответствующие разным собственным числам ϱ_i^2 и ϱ_j^2. В системе

$$\begin{cases} (A - \varrho_i^2 \Sigma_{22}) M_i = 0 \\ (A - \varrho_j^2 \Sigma_{22}) M_j = 0 \end{cases}$$

первое уравнение умножим на M_j^T, второе на M_i^T, и уравнения сложим:

$$\begin{cases} M_j^T A M_i - \varrho_i^2 M_j^T \Sigma_{22} M_i = 0 \\ M_i^T A M_j - \varrho_j^2 M_i^T \Sigma_{22} M_j = 0 \end{cases} \implies (\varrho_i^2 - \varrho_j^2) M_i^T \Sigma_{22} M_j = 0,$$

откуда $cov(V_{2i}, V_{2j}) = EV_{2i}V_{2j} = EM_i^T U_2 U_2^T M_j = M_i^T \Sigma_{22} M_j = 0$.

Аналогично $cov(V_{1i}, V_{1j}) = EV_{1i}V_{1j} = EL_i^T U_1 U_1^T L_j = L_i^T \Sigma_{11} L_j = 0$.

$$cov(V_{1i}, V_{2j}) = EL_1^T U_1 U_2^T M_j = L_i^T \Sigma_{12} M_j$$

Рассмотрим первое уравнение из системы (93) в случае $\varrho_i \neq 0$ и умножим его обе части на L_i^T:

$$\Sigma_{12}M_i = \varrho_i\Sigma_{11}L_i \iff L_i^T\Sigma_{12}M_i = \varrho_i L_i^T\Sigma_{11}L_i = \varrho_i,$$

при умножении на L_j^T получим $L_j^T\Sigma_{12}M_i = \varrho_i L_j^T\Sigma_{11}L_i = 0$. Если $\varrho_i = 0$, тогда $\Sigma_{12}M_i = 0 \iff L_i^T\Sigma_{12}M_i = 0$.

Для проверки значимости равенства нулю канонических корреляций используются статистики хи-квадрат.

9.4. Пример канонических корреляций

Исследуем корреляционную структуру между психологическими признаками и показателями тяжести наркотической зависимости.

Индекс тяжести	психологические характеристики
X_1 – медицинский статус	Y_1 – рискованность
X_2 – наркотический статус	Y_2 – уровень депрессии
X_3 – проблемы с законом	Y_3 – уровень тревоги
X_4 – психиатрический статус	

Первая каноническая корреляция между каноническими величинами вида:

$$U_1 = -0.45X_1 - 0.45X_2 + 0.31X_3 - 0.56X_4;$$
$$V_1 = -0.07Y_1 - 0.31Y_2 - 0.80Y_3$$

равна $R_1 = 0.352$ ($p = 0.000001$). Она указывает на корреляцию между уровнем тревоги и более тяжелым медицинским статусом, связанным с тяжестью наркотической зависимости и психиатрическим статусом.

Канонические факторные нагрузки

	$R1 = 0.352$	$R_2 = 0.193$	$R_3 = 0.143$
мед.ст.	-0.641	0.298	-0.707
нарк.ст.	-0.545	0.199	0.390
проблемы с законом	0.296	-0.628	-0.227
психиатр. ст.	-0.663	-0.704	0.111
рискованность	-0.091	-0.934	0.346
депрессия	-0.735	-0.243	-0.633
тревожность	-0.961	0.182	0.209

Вторая каноническая корреляция между каноническими величинами вида

$$U_2 = 0.24X_1 + 0.28X_2 - 0.56X_3 - 0.74X_4;$$
$$V_2 = -0.90Y_1 - 0.37Y_2 + 0.36Y_3,$$

равная $R_2 = 0.193$ ($p = 0.014$), указывает на взаимосвязь между рискованностью и тяжестью наркотической зависимости в форме проявления психиатрического статуса и проблем с законом.

Третья каноническая корреляция между каноническими величинами вида

$$U_3 = -0.91X_1 + 0.57X_2 - 0.50X_3 + 0.21X_4;$$
$$V_3 = 0.43Y_1 - 1.08Y_2 + 0.79Y_3$$

равная $R_3 = 0.143$ ($p = 0.06$), указывает на взаимосвязь между депрессией и более тяжелым медицинским статусом.

В целом, выделяются три общих фактора: непосредственно тяжесть наркотической зависимости, психиатрическое отягощение и состояние организма. Первый сопровождается повышением тревожности с депрессией, второй рискованностью, а третий рафинированной депрессией.

Информация о величине и значимости канонических коэффициентов корреляции находится во вкладке *Canonical factos- Chi square tests*, факторные нагрузки *Factor structures - Factor structures*

and redundancies, канонические коэффициенты в *Canonical scores - Left and right set canonical weights.*

10. Анализ данных типа времени жизни

10.1. Кривая дожития

В анализе данных типа времени жизни исследуются объекты, для которых определено точечное событие, которое условно называют "отказ". Об этом событии известно, что оно происходит один раз и через некоторое случайное время $\tau > 0$. Например, 1) "отказ" – это наступление рефлекса выведения молока, τ – время до его наступления; 2) "отказ" – это инфаркт, или инсульт, или смерть, τ – временной интервал до его наступления.

Исходные данные представляют собой времена ожидания отказа $\tau_i, i = 1, \ldots, n$ для каждого из индивидов. Рассмотрим $n = 22$ наблюдения τ – времени ремиссии (в неделях).

τ_i :	1	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11	12	12	15	17	22	23

По определению, кривая дожития – это $F(t) = P\{\tau > t\}$ – вероятность того, что отказ наступит после момента времени t. Обозначим через d_t количество отказов в момент времени t, r_t – количество объектов, у которых отказ наступает не раньше момента t, считаем, что $r_1 = n$. Тогда при полном наличии информации о временах отказа оценка кривой дожития вычисляется таким образом:

$$\widehat{F}_t = \frac{r_t - d_t}{n} = \frac{r_{t+1}}{n}\,. \tag{94}$$

t	r_t	d_t	$r_t - d_t$	$\widehat{F}_t$
0	22	0	22	1
1	22	3	19	$\frac{19}{22}$
2	19	2	17	$\frac{17}{22}$
3	17	1	16	$\frac{16}{22}$

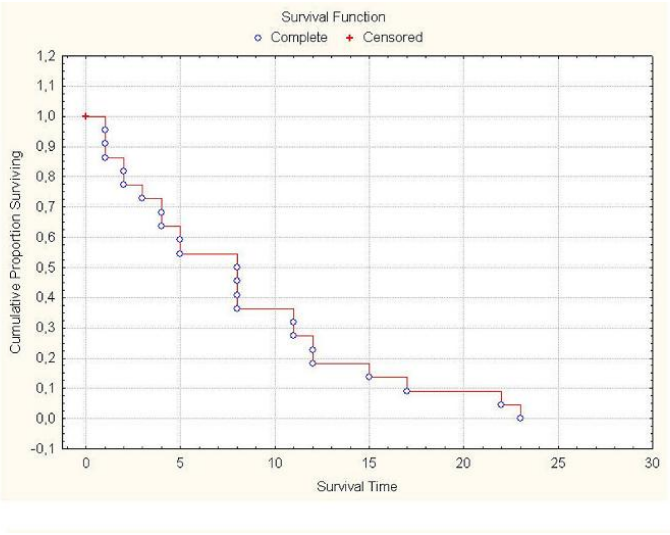


Рис. 11. Кривая дожития по времени ремиссии (без цензурирования).

Для фиксации данных требуется:

- четко установить начало отсчета,

- определить масштаб времени,

- определить понятие отказа.

Чаще всего стоит задача оценки вероятности дожития или сравниваются времена ожидания отказов в двух или нескольких выборках.

10.2. Функция риска

Кривую дожития иногда называют функцией надежности. Ее производная с обратным знаком называется плотностью и обозначается через

$f(t)$.

$$f(t) = -F'(t) = -\lim_{\Delta\to 0}\frac{F(t+\Delta)-F(t)}{\Delta} = \lim_{\Delta\to 0}\frac{F(t)-F(t+\Delta)}{\Delta} =$$
$$= \lim_{\Delta\to 0}\frac{P\{\tau > t\} - P\{\tau > t+\Delta\}}{\Delta} = \lim_{\Delta\to 0}\frac{P\{t < \tau \le t+\Delta\}}{\Delta}$$

По смыслу, плотность $f(t)$ равна вероятности гибели в момент t.

Условная вероятность гибели в момент времени t выражается функцией интенсивности $h(t)$ вида

$$h(t) = \lim_{\Delta\to 0}\frac{P\{t < \tau \le t+\Delta | \tau > t\}}{\Delta} = \lim_{\Delta\to 0}\frac{P\{t < \tau \le t+\Delta\}}{\Delta P\{\tau > t\}} = \frac{f(t)}{F(t)}, \quad (95)$$

$$\text{или} \quad h(t) = \frac{-F'(t)}{F(t)} = (-\ln F(t))'.$$

Минус логарифма от функции надежности называется *функцией риска:*

$$H(t) = -\ln F(t) \qquad F(t) = e^{-H(t)}.$$

Поскольку $F(0) = 1$, то $H(0) = 0$. Тогда учитывая, что $H'(t) = h(t)$, получаем:

$$H(t) = H(t) - H(0) = \int_0^t h(u)du \quad \text{и} \quad F(t) = e^{-\int_0^t h(u)du}.$$

Если риск (интенсивность) является постоянным $h(u) = \varrho$, то функция надежности убывает по экспоненте: $F(t) = e^{-\varrho t}$.

10.3. Цензурирование и оценка Каплана-Мейера

Цензурирование – это неполное наблюдение за временем ожидания отказа. Некоторые объекты не могут наблюдаться в течение полного времени до отказа. Известным оказывается только то, что до некоторого времени t отказ еще не наступил. Например, больной состоял на медицинском учете,

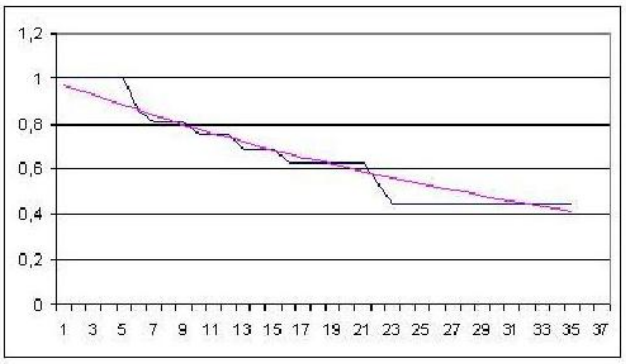


Рис. 12. Кривые дожития времени ремиссии с цензурированием и ее экспоненциальная модель).

а потом уехал в другой город, и о нем больше ничего не известно, или у животного фиксируемый рефлекс не наступил за установленное время эксперимента. При этом наблюдения записываются в виде пары признаков: в одном указывается или время наступления отказа, или время цензурирования, второй признак является индикатором цензурирования.

Если обозначить через τ_i время ожидания отказа, c_i – момент цензурирования, то исходные данные можно представить в виде:

$$(X_i, V_i), \quad \text{где } X_i = \min(\tau_i, c_i), \quad i = 1, \ldots, n,$$
$$V_i = \begin{cases} 0, & \text{если } \tau_i \le c_i \quad (\text{отказ }), \\ 1, & \text{если } \tau_i > c_i \quad (\text{цензурирование}). \end{cases}$$

При наличии цензурирования в оценке кривой дожития используется другой принцип. Обозначим через p_i вероятность того, что при условии дожить до момента i отказ в этот момент так и не наступил. В качестве оценки вероятности p_i будем рассматривать отношение

$$\widehat{p}_i = \frac{r_i - d_i}{r_i} = 1 - \frac{d_i}{r_i}, \tag{96}$$

где r_i число наблюдений с отказом не ранее момента i, d_i – число отказов

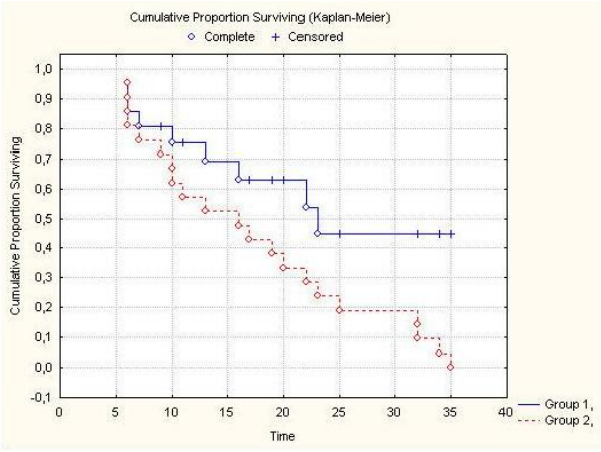


Рис. 13. Кривые дожития времени ремиссии с цензурированием и без цензурирования.

в момент i. Тогда

$$P\{\tau > t\} = P\{\tau > t-1\}P\{\tau > t|\tau > t-1\} = P\{\tau > t-1\}p_t$$

равна произведению вероятностей $P\{\tau > t\} = p_1 p_2 \dots p_t$, так как $P\{\tau > 0\} = 1$. Следовательно,

$$\widehat{P}\{\tau > t\} = \frac{r_1 - d_1}{r_1}\frac{r_2 - d_2}{r_2}\dots\frac{r_t - d_t}{r_t} \tag{97}$$

Обозначим через $h_i = \frac{d_i}{r_i}$ функцию интенсивности или условную вероятность погибнуть в момент i. Оценка кривой дожития, выраженная через интенсивности, носит название функции дожития Каплана-Мейера:

$$\widehat{P}\{\tau > t\} = \prod_{i=1}^{t}(1 - h_i)\ . \tag{98}$$

Для построения функции Каплана-Мейер рассмотрим данные с цензурированием, представленные в виде пары признаков:

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
X_j	6	6	6	6	7	9	10	10	11	13	16	17	19	20	22	23	25	32	32	34	35
V_j	1	0	0	0	0	1	1	0	1	0	0	1	1	1	0	0	1	1	1	1	1

i	d_i отказы	z_i цензурирование	$r_i = r_{i-1} -$ $-d_{i-1} - z_{i-1}$ оставшиеся	$\widehat{p}_i = 1 - h_i$ усл.вер-ть	$\widehat{F}_i$ $\prod_{j=1}^{i}(1-h_j)$
1	0	0	21	1	1
2	0	0	21	1	1
3	0	0	21	1	1
4	0	0	21	1	1
5	0	0	21	1	1
6	3	1	21	0.857	0.857
7	1	0	17	0.941	0.807
8	0	0	16	1	0.807
9	0	1	16	1	0.807
10	1	1	15	0.933	0.753
11	0	1	13	1	0.753
12	0	0	12	1	0.753
13	1	0	12	0.917	0.690
14	0	0	11	1	0.690
15	0	0	11	1	0.690
16	1	0	11	0.909	0.627
17	0	1	10	1	0.627
18	0	0	9	1	0.627
19	0	1	9	1	0.627
20	0	1	8	1	0.627
21	0	0	7	1	0.627
22	1	0	7	0.857	0.538
23	1	0	6	0.833	0.448
24	0	0	5	1	0.448
25	0	1	5	1	0.448
26	0	0	4	1	0.448
27	0	0	4	1	0.448
28	0	0	4	1	0.448
29	0	0	4	1	0.448
30	0	0	4	1	0.448
31	0	0	4	1	0.448
32	0	2	4	1	0.448
33	0	0	2	1	0.448
34	0	1	2	1	0.448
35	0	1	1	1	0.448

При отсутствии цензурирования количество выживших в момент $i+1$ равно $r_{i+1} = r_i - d_i$, поэтому в формуле (98) элементы в соседних дробях сокращаются, и остается выражение

$$\widehat{P}\{\tau > t\} = \frac{r_t - d_t}{r_1} = \frac{r_{t+1}}{r_1},$$

соответствующее оценке (94).

Если в интенсивности $h_i = d_i/r_i$ числитель и знаменатель разделить на n, то получим выражение интенсивности как отношение вероятности погибнуть в момент i к вероятности дожить до этого момента (95):

$$h_i = \frac{d_i/n}{r_i/n} = \frac{\widehat{f}_i}{\widehat{F}_{i-1}} = \frac{\hat{f}_i}{\hat{f}_i + \hat{f}_{i+1} + \ldots}.$$

По аналогии для генеральных дискретных характеристик получаем

$$h_1 = f_1, \quad \text{так как} \quad h_1 = \frac{f_1}{f_1 + f_2 + \ldots}, \quad \Longrightarrow \quad F_0 = \frac{f_1}{h_1} = 1;$$

$$h_2 = \frac{f_2}{f_2 + f_3 + \ldots} = \frac{f_2}{1 - f_1} = \frac{f_2}{1 - h_1}, \quad \Longrightarrow \quad f_2 = h_2(1 - h_1),$$

$$\Longrightarrow \quad F_1 = \frac{f_2}{h_2} = (1 - h_1),$$

$$f_3 + f_4 + \ldots = 1 - f_1 - f_2 = 1 - h_1 - h_2(1 - h_1) = (1 - h_1)(1 - h_2),$$

$$\Longrightarrow \quad f_3 = h_3(1 - h_1)(1 - h_2), \quad F_2 = (1 - h_1)(1 - h_2),$$

далее аналогично.

В *StatSoft* анализ данных типа времени жизни осуществляется при помощи программы *Statistics - Advanced Linear - Nonlinear Models - Survivel Analysis*.

Выбираем *Kaplan and Meier product – limit method*. Данные должны быть представлены по крайней мере в двух столбцах. В одном из них должны быть указаны наработки до отказа или до цензурирования, обозначим его, например, через X, во втором – индикатор C цензурирования. В *Variables* указываем требуемые переменные. После активизации признака C появляются окна, в которых нужно указать коды цензурирования. Для получения графика, подобного рис.11, нужно во вкладке *Kaplan – Meier plots* выбрать *Survivel times vs.cum. proption surviving*.

10.4. Оценка параметра постоянного риска

Функцией правдоподобия $L(x_1, \ldots, x_n)$ называется вероятность появления данной выборки $x_i = \min(t_i, c_i),\ i = 1, \ldots, n$.

$$L(x_1, \ldots, x_n) = \prod_{t_i} f(t_i, \varrho) \prod_{c_i} F(c_i, \varrho)\,,$$
$$\ln L(x_1, \ldots, x_n) = \sum_{t_i} \ln f(t_i, \varrho) + \sum_{c_i} \ln F(c_i, \varrho)\,.$$

Поскольку $f(t) = h(t)F(t)$, то

$$\ln L(x_1, \ldots, x_n) = \sum_{t_i} \ln h(t_i, \varrho) + \sum_{t_i} \ln F(t_i, \varrho) + \sum_{c_i} \ln F(c_i, \varrho) \ =$$

$$= \sum_{t_i} \ln h(t_i, \varrho) + \sum_{x_i} \ln F(x_i, \varrho) = \sum_{t_i} \ln h(t_i, \varrho) - \sum_{x_i} H(x_i, \varrho)\,.$$

Если риск постоянный, то $h(t_i, \varrho) = \varrho$, $H(t) = \varrho t$. Следовательно,

$$\ln L(x_1, \ldots, x_n) = \ln \varrho d - \varrho \sum_{x_i} t\,,$$

где d означает общее число отказов, $\sum\limits_{x_i} t$ – суммарную наработку. Оценки максимального правдоподобия находятся дифференцированием функции правдоподобия $\ln L(x_1, \ldots, x_n)$ по параметру ϱ.

$$\frac{\partial \ln L}{\partial \varrho} = \frac{d}{\varrho} - \sum_{x_i} t = 0.$$

Таким образом, параметр экспоненциальной модели кривой дожития вычисляется как отношения числа отказов к суммарной наработке. В нашем примере $\widehat{\varrho} = \frac{9}{359} = 0.025$.

При отсутствии цензурирования $\frac{2d\varrho}{\hat{\varrho}}$ имеет распределение хи-квадрат с $2d$ степенями свободы. Действительно, суммарная наработка имеет гамма

распределения

$$\sum_{i=1}^{d} x_i \sim \gamma(\varrho, d), \quad \frac{1}{d}\sum_{i=1}^{d} x_i \sim \gamma(d\varrho, d),$$

$$2d\varrho\left(\frac{1}{d}\sum_{i=1}^{d} x_i\right) = \frac{2d\varrho}{\hat{\varrho}} \sim \gamma(d\varrho\left(\frac{1}{2d\varrho}\right), d) = \gamma\left(\frac{1}{2}, \frac{2d}{d}\right) = \chi^2(2d)\,.$$

10.5. Доверительные интервалы для кривых дожития

При незначительном влиянии цензурированных данных используется формула для вычисления дисперсии

$$\mathbf{D}(\hat{F}_j) = \frac{\hat{F}_j^2(1-\hat{F}_j)}{r_{j+1}}\,.$$

Пусть n_0 число нецензурированных данных, $\hat{F}_j = \frac{r_{j+1}}{n_0}$, значение $n_0 = \frac{r_{j+1}}{\hat{F}_j}$ подставляем в выражение для дисперсии биномиальной частоты $\hat{F}_j$ в n_0 испытаниях:

$$\mathbf{D}(\hat{F}_j) = \frac{\hat{F}_j(1-\hat{F}_j)}{n_0} = \frac{\hat{F}_j^2(1-\hat{F}_j)}{r_{j+1}}\,.$$

Обозначим через $s_j = r_j - d_j$ число выживших при $t = j$. Если p_j - вероятность выжить в этот момент, то s_j имеет биномиальное распределение $\mathcal{B}(p_j, r_j)$. Формула Гринвуда:

$$\mathbf{D}(\hat{F}_j) = \hat{F}_j^2 \sum_{j=1}^{t} \frac{d_j}{r_j(r_j - d_j)}$$

Обозначим через $F_j = \mathbf{E}\hat{F}_j$, где $\hat{F}_j = \prod\limits_{j=1}^{t} \frac{s_j}{r_j}$. Очевидно, что $\mathbf{E}\left(\frac{s_j}{r_j}\right) = p_j$,

$$\frac{\hat{F}_j}{F_j} = \prod_{j=1}^{t} \frac{s_j}{r_j p_j} = \prod_{j=1}^{t}\left(1 + \frac{s_j - r_j p_j}{r_j p_j}\right) \approx 1 + \sum_{j=1}^{t} \frac{s_j - r_j p_j}{r_j p_j}$$

$$\mathbf{D}\left(\frac{\hat{F}_j}{F_j}\right) = \sum_{j=1}^{t} \frac{r_j p_j (1-p_j)}{(r_j p_j)^2} = \sum_{j=1}^{t} \frac{1-p_j}{r_j p_j} \approx \sum_{j=1}^{t} \frac{1-\hat{p}_j}{r_j \hat{p}_j},$$

$$\mathbf{D}\hat{F}_j = {F_j}^2 \sum_{j=1}^{t} \frac{1-\hat{p}_j}{r_j \hat{p}_j} \approx \hat{F}_j^2 \sum_{j=1}^{t} \frac{1-\hat{p}_j}{r_j \hat{p}_j} = \hat{F}_j^2 \sum_{j=1}^{t} \frac{d_j}{r_j(r_j - d_j)}\,.$$

10.6. Регрессионная модель Кокса

Рассматривается модель пропорциональных интенсивностей

$$h(t,z) = \psi(z,\beta) h_0(t) = e^{\beta^T z} h_0(t)\,,$$

в которой для любого объекта вектор поясняющих переменных z не зависит от времени. Будем считать, что базовая функция интенсивности $h_0(t)$ не известна. Обозначим через $\tau_1 < \tau_2 < \ldots < \tau_n$ упорядоченные моменты наработок до отказов у n объектов, T_j – номер объекта, отказавшего в момент τ_j, то есть

$$T_j = i \;\Leftrightarrow\; t_i = \tau_j\,.$$

Введем обозначение для множества риска – множества объектов, действующего непосредственно перед j-м отказом: $R(\tau_j) = \{i : t_i \geq \tau_j\}$, r_j – размер этого множества.

Пусть известна полная история до момента τ_j

$$H_{j-1} = \{\tau_1, \tau_2, \ldots, \tau_{j-1}; i_1, i_2, \ldots, i_{j-1}\}\,,$$

и один из объектов множества риска $R(\tau_j)$ отказал в момент τ_j. Условная вероятность того, что это был i-й объект, равна

$$p_j(i|i_1, i_2, \ldots, i_{j-1}) = \frac{h_i(\tau_j)}{\sum\limits_{k \in R(\tau_j)} h_k(\tau_j)} = \frac{\psi(i)}{\sum\limits_{k \in R(\tau_j)} \psi(k)}\,,$$

где $\psi(i) = \psi(z_i, \beta)$ есть множитель, влияющий на интенсивность i-го объекта. По цепному правилу для условных вероятностей получаем выражение для функции правдоподобия или вероятность истории H_n

$$p(i_1, \ldots, i_n) = \prod_{j=1}^{n} p_j(i_j | i_1, i_2, \ldots, i_{j-1}) = \prod_{j=1}^{n} \frac{\psi(i_j)}{\sum\limits_{k \in R(\tau_j)} \psi(k)} .$$

При наличии цензурирования произведение осуществляется по отказам:

$$lik = \prod_{i \in D} \frac{\psi(i)}{\sum\limits_{k \in R_i} \psi(k)},$$

где через D обозначено множество отказавших объектов, через R_i соответствующее i множество риска. При $\psi(z; \beta) = e^{\beta^T z}$ для оценивания параметров вектора β логарифм функции правдоподобия

$$\ln(lik) = \sum_{i \in D} \left(\ln \psi(i) - \ln \sum_{k \in R(\tau_j)} \psi(k) \right) = \sum_{i \in D} \left(\beta^T z_i - \ln \sum_{k \in R(\tau_j)} e^{\beta^T z_k} \right)$$

дифференцируют по параметрам, приравнивают производные к нулю и решают систему уравнений:

$$\sum_{i \in D} (z_{i\nu} - A_{i\nu}(\beta)) = 0, \ \nu = 1, 2, \ldots, K,$$

где $z_{i\nu}$ – значение ν-й из K компонент поясняющей переменной z для i-го объекта,

$$A_{i\nu}(\beta) = \frac{\sum\limits_{k \in R_i} z_{k\nu} e^{\beta^T z_k}}{\sum\limits_{k \in R_i} e^{\beta^T z_k}}$$

взвешенное среднее переменной $z_{i\nu}$ по соответствующему множеству риска.

В случае двух выборок имеется одна $(K = 1)$ поясняющая переменная z, которая принимает два значения: 0, если объект принадлежит нулевой выборке, и 1, если принадлежит первой выборке.

$$\psi(i) = \begin{cases} e^{\beta}, & \text{i-ый объект из группы 1} \\ 1, & \text{i-ый объект из группы 0} \end{cases}$$

Обозначим через d_1 количество отказов в первой выборке, через r_{0i} и r_{1i} объемы множеств риска для каждой выборки в момент i. Логарифм функции правдоподобия равен

$$\ln(lik) = \sum_{i\in D}\left(\ln\psi(i) - \ln(\sum_{k\in R_i}\psi(k))\right) = \beta d_1 - \sum_{i\in D}\ln(r_{1i}e^{\beta} + r_{0i}),$$

$$\frac{d\ln(lik)}{d\beta} = d_1 - \sum_{i\in D}\frac{r_{1i}e^{\beta}}{r_{1i}e^{\beta} + r_{0i}} = 0.$$

Из последнего уравнения находится оценка параметра $\widehat{\beta}$.

10.7. Сравнение двух групп по кривым дожития

Статистики, используемые для проверки гипотезы однородности, можно получить из одной таблицы. Для упрощения рассмотрим нецензурированные данные. Обозначим через d_{0i} и d_{1i} количество отказов в группе 0 и 1 в момент времени i, через $d_i = d_{0i} + d_{1i}$ число отказов в обеих группах. Пусть r_{0i} и r_{1i} – количество доживших до момента i.

время	группа 0		группа 1		всего		*Peto Cox*		Cox-Mantel	Log-rank
i	r_{0i}	d_{0i}	r_{1i}	d_{1i}	r_i	d_i	$\frac{d_i r_{0i}}{r_i}$	$\frac{d_i r_{1i}}{r_i}$	$\frac{d_i r_{0i} r_{1i}}{r_i^2}$	$\frac{d_i r_{0i} r_{1i}(r_i - d_i)}{r_i^2(r_i-1)}$
1	3	1	3	0	6	1	0.5	0.5	0.25	0.25
2	2	1	3	0	5	1	0.4	0.6	0.24	0.24
3	1	0	3	1	4	1	0.25	0.75	0.1875	0.1875
4	1	0	2	1	3	1	$\frac{1}{3}$	$\frac{2}{3}$	0.22	0.25
5	1	1	1	0	2	1	0.5	0.5	0	1.15
6	0	0	1	1	1	1	0	1		
сумма		3		3			1.98	4.02	1.15	1.15
		d_0		d_1			E_0	E_1	D_{cox}	D

Обозначим через E_0 и E_1 математические ожидания числа отказов

$$d_0 = \sum_{i=1}^{n} d_{0i}, \quad d_1 = \sum_{i=1}^{n} d_{1i}.$$

При $\beta = 0$ ожидаемые числа отказов как суммы математических ожиданий

гипергеометрических случайных величин равны

$$\widehat{E}_0 = \sum_{i=1}^{n} \frac{d_i r_{0i}}{r_i}, \quad \widehat{E}_1 = \sum_{i=1}^{n} \frac{d_i r_{1i}}{r_i}\,.$$

В каждый момент времени общее множество рисков имеет мощность $r_i = r_{0i} + r_{1i}$, всего отказов d_i. Если ξ случайное число отказов из 0-группы подчиняется гипергеометрическому распределению:

$$P\{\xi = d_{0i}\} = \frac{C_{r_{0i}}^{d_{0i}} C_{r_{1i}}^{d_{1i}}}{C_{r_i}^{d_i}}$$

с математическим ожиданием $\mathbf{E}\xi = \frac{d_i r_{0i}}{r_i}$ и дисперсией $\mathbf{D}\xi = d_i \frac{r_{0i} r_{1i}(r_i - d_i)}{r_i^2(r_i - 1)}$. Лог-ранговый критерий проверки однородности основан на асимптотической нормальности числа отказов d_0, которое в случае однородных выборок имеет нормальное распределение со средним E_0 и дисперсией

$$D = \sum_{i=1}^{n} d_i \frac{r_{0i} r_{1i}(r_i - d_i)}{r_i^2(r_i - 1)}.$$

Статистика лог-рангового критерия вида

$$\chi^2 = \frac{(d_0 - E_0)^2}{D} = 0.904,\ p = 0.34$$

в случае однородных кривых дожития имеет распределение хи-квадрат с одной степенью свободы $\chi^2(1)$.

Пусть $D_{cox} = \sum\limits_{i=1}^{n} \frac{d_i r_{0i} r_{1i}}{r_i^2}$. Первая приближенная статистика Cox-Mantel вида:

$$\chi^2 = \frac{(d_0 - E_0)^2}{D_{cox}} = 0.904,\ p = 0.34$$

также имеет в случае однородных кривых дожития распределение $\chi^2(1)$. Вторая приближенная статистика Peto вида:

$$\chi^2 = \frac{(d_0 - E_0)^2}{E_0} + \frac{(d_1 - E_1)^2}{E_1} = 0.784\,,\ p = 0.376,$$

имеет в случае однородных кривых дожития распределение $\chi^2(1)$. При построении статистики Кокса используется отношение

$$F = \frac{\widehat{E}_0/(2d_0)}{\widehat{E}_1/(2d_1)} = 2.03,\ p = 0.2,$$

которое имеет распределение Фишера с $2d_0$ и $2d_1$ степенями свободы.

Рис. 14. Сравнение кривых дожития.

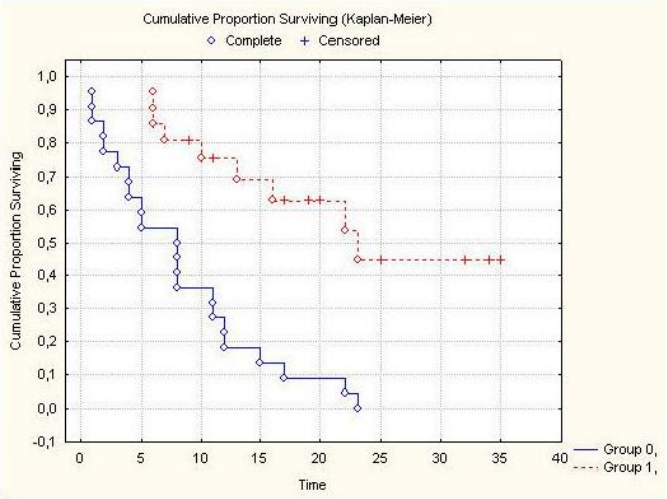


Рис. 15. Кривые дожития по времени ремиссии для двух групп.

11. Кластерный анализ по категориальным признакам

11.1. Информационное разнообразие групп и информационный выигрыш

Качественными или категориальными называются признаки, градации которых можно менять местами. Например, цветы могут быть красными, синими, белыми, желтыми или лилиями, розами, гиацинтами и т.д. Пусть имеется группа n индивидов (букет цветов), характеризуемая признаком ξ (тип цветка) со значениями $x_1, \ldots, x_m$ (ромашка, колокольчик, гвоздика и т.д.), $n = a_1 + a_2 + \ldots + a_m$, где a_i – количество индивидов со значением x_i признака ξ (a_1 число ромашек, a_2 колокольчиков и так далее).

Информационное разнообразие I этой группы измеряется формулой:

$$I = n \ln n - \sum_{i=1}^{m} a_i \ln a_i.$$

Например, информационное разнообразие букета, состоящего из двух гвоздик, одного тюльпана и двух хризантем равно

$$I_1 = 5 \ln 5 - 2 \ln 2 - 1 \ln 1 - 2 \ln 2 = 5.275.$$

В группе с одинаковыми значениями признака информационное разнообразие равно нулю, так как $I = n \ln n - 0 \cdot \ln 0 - n \ln n = 0$. Очевидно, что разнообразие группы, состоящего из одного индивида, также равно нулю.

Если группа индивидов характеризуется несколькими признаками (например, в букете гвоздики и тюльпан красные, а хризантемы белые), то разнообразие складывается из информационных разнообразий по каждому признаку. Разнообразие по цвету равно:

$$I_2 = 5 \ln 5 - 3 \ln 3 - 2 \ln 2 = 3.365,$$

отсюда общее разнообразие равно $I = I_1 + I_2 = 5.275 + 3.365 = 8.64$.

Информационный выигрыш от слияния двух групп A и B с разнообразиями I_A и I_B равен разности между разнообразием объединенной группы и разнообразиями исходных групп до объединения:

$$\Delta I_{A+B} = I_{A+B} - I_A - I_B.$$

Информационный выигрыш от слияния групп может быть нулевым или положительным. Например, если объединяются букет роз и букет ромашек, разнообразие которых по-отдельности равно нулю, то выигрыш положительный, так как разнообразие смешанного букета положительно. Если к ромашке с колокольчиком присоединить еще одну пару ромашки с колокольчиком, то информационный выигрыш окажется нулевым, так как

$$\begin{aligned} I_A &= 2\ln 2 - 1\ln 1 - 1\ln 1 = 2\ln 2, \\ I_B &= 2\ln 2, \\ I_{A+B} &= 4\ln 4 - 2\ln 2 - 2\ln 2 = 4\ln 4 - 4\ln 2 = 4(\ln 4 - \ln 2) = 4\ln 2, \\ \Delta I &= 4\ln 2 - 2\ln 2 - 2\ln 2 = 0. \end{aligned}$$

11.2. Процедура кластеризации индивидов

Рассмотрим процедуру построения кластеров для шести лиц пяти типов, условно изображенных на рис.16. Каждый индивид характеризуется четырьмя признаками глаза: 1 - большие, круглые, 0 - узкие, мелкие; нос: 1 - большой, толстый, 0 - тонкий, маленький; рот: 1 - большой, 0 - узкий; уши: 1 - большие, музыкальные, 0 - маленькие. Первый индивид характеризуется вектором признаков (1110), второй (1010), два третьих (1101), четвертый (0101) и пятый (0000). Сначала для каждой пары индивидов вычисляем количество признаков, по которым индивиды отличаются,

	1	2	3	4	5
1	–	1	2	3	3
2		–	3	4	2
3			–	1	3
4				–	2
5					–

и получаем, что больше всего друг на друга походят индивиды первый и второй (отличаются только носами) , а также два третьих и четвертый (только глазками). Таким образом, мы агрегируем этих индивидов в три группы: $A(1,2)$, $B(3,3,4)$ и $C(5)$. Соответственно их разнообразия равны

$$I_{A(1,2)} = 2\ln 2 - 1\ln 1 - 1\ln 1 = 1.37,$$
$$I_{B(3,3,4)} = 3\ln 3 - 2\ln 2 - 1\ln 1 = 1.91,$$
$$I_C = 0.$$

В табл.5 показано, как вычисляются эти разнообразия и разнообразия от объединения групп. Отсюда информационные выигрыши от объединения групп соответственно равны:

$$\Delta I_{A+B} = I_{A+B} - I_A - I_B = 11.73 - 1.37 - 1.91 = 8.44,$$
$$\Delta I_{A+C} = I_{A+C} - I_A - I_C = 5.73 - 1.37 - 0 = 4.34,$$
$$\Delta I_{B+C} = I_{B+C} - I_B - I_C = 7.27 - 1.91 - 0 = 5.36.$$

Наименьший выигрыш достигается от объединения групп $A(1,2)$ и $C(5)$. Поэтому получаем два кластера: (1, 2, 5) и (3,3,4), которые отличаются ушками (рис.16).

Задание 1. Рассмотрим кластеры $F(1,2,5)$ и $B(3,3,4)$. Нужно заполнить таблицу 6, для этого не обязательно пересчитывать, можно воспользоваться результатами табл.5, затем вычислить информационные выигрыши

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{A+C}	1.91	1.91	1.91	0	5.73

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{B+C}	2.77	2.25	0	2.25	7.27

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
		I_{A+B}	2.5	2.5	3.37	3.37	11.73

Таблица 5. Вычисление разнообразия объединений групп $A(1,2)$-$C(5)$, $B(3,3,4)$-$C(5)$ и $A(1,2)$-$B(3,3,4)$.

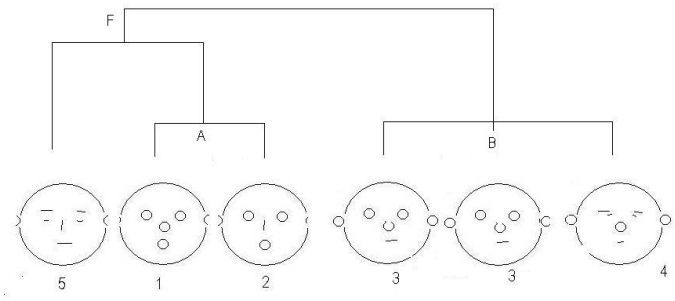


Рис. 16. Дерево кластеризации.

от объединения

$$\Delta I_{F+D} = I_{F+D} - I_F - I_D,$$
$$\Delta I_{B+D} = I_{B+D} - I_B - I_D,$$

и по наименьшему из этих выигрышей выяснить, к кому из них примкнет $D(6)$ индивид (1001) шестого типа (рис.17).

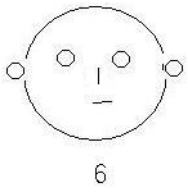


Рис. 17. Индивид (1001) шестого типа.

11.3. Кластеризация индивидов в случае метрических признаков

Пусть имеется n индивидов и p признаков, x_{ij} – наблюдение i-го признака и j-го индивида. В качестве мер сходства и различия используются следующие метрики:

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
	2		1	0	1	0	
F	5		0	0	0	0	
		I_F					
D	6		1	0	0	1	
		I_D					
		I_{F+D}					

Таблица 6. Форма для вычисления разнообразия объединений групп $F(1,2,5)$-$D(6)$, $B(3,3,4)$-$D(6)$.

1. евклидова метрика, равная геометрическому расстоянию между точками (например, индивидами 1 и 2) в p-мерном пространстве:

$$\sqrt{\sum_{k=1}^{p}(x_{i1}-x_{i2})^2}$$

2. квадрат евклидовой метрики:

$$\sum_{k=1}^{p}(x_{i1}-x_{i2})^2$$

3. манхеттенская метрика:

$$\sum_{k=1}^{p}|x_{i1}-x_{i2}|$$

4. Метрика Брея-Картиса:

$$\frac{\sum\limits_{k=1}^{p}|x_{i1}-x_{i2}|}{\sum\limits_{k=1}^{p}x_{i1}+\sum\limits_{k=1}^{p}x_{i2}}$$

5. Канберровская метрика:

$$\sum_{k=1}^{p}\left(\frac{|x_{i1}-x_{i2}|}{x_{i1}+x_{i2}}\right).$$

6. Метрика Адкинса:

$$\sum_{k=1}^{p}\left(\frac{|x_{i1}-x_{i2}|}{|x_{i1}|+|x_{i2}|}\right).$$

Последние три метрики используются часто при большом различии масштабов измерения. Например, если единицы измерения одного признака миллиметры, а другого километры, то ненормированные евклидовы и манхеттенские метрики будут измерять меру сходства и различия только по признаку с большими единицами измерения.

Кластеризация индивидов осуществляется на основе выбора специальной стратегии объединения. Сначала рассматриваются всевозможные пары индивидов и пары с минимальным различием объединяются в группы. Далее встает вопрос, как измерить расстояние между группами. Существует несколько стратегий измерения расстояния между группами. Все они по формальному признаку сводятся к выражению:

$$d_{kh} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|,$$

где через h обозначена группа, полученная в результате объединения индивидов i и j, $h = i \cup j$, через k – группа, расстояние до которой от группы h определяется в результате выбранной стратегии. Параметры $\alpha_i, \alpha_j, \beta$ и γ называются параметрами стратегии.

Например, параметры вида $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = -0.5$ определяют стратегию ближнего соседа, когда за расстояние между группами принимается расстояние между ближайшими соседями. Эта стратегия уменьшает расстояния между группами или, как говорят, сжимает пространство.

Параметры вида $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = 0.5$ определяют стратегию дальнего соседа, когда за расстояние между группами принимается самое дальнее расстояние между элементами групп. Эта стратегия расширяет пространство.

Параметры $\alpha_i = n_i/n_h, \alpha_j = n_j/n_h$, $\beta = 0$, $\gamma = 0$ определяют стратегию группового среднего, когда за расстояние между группами принимается расстояние между ближайшими соседями. Через n_i обозначено количество элементов в i-й группе. Недостаток этой стратегии – немонотонность, но она не изменяет пространство.

Существует еще центроидная стратегия, когда расстояние между группами определяется как расстояние между центрами тяжести.

Стратегия, удовлетворяющая условиям монотонности и сохранения метрики пространства, называется гибкой стратегией. Соответствующие ей параметры имеют вид: $\alpha_i = \alpha_j = 0.625$, $\beta = -0.25$, $\gamma = 0$.

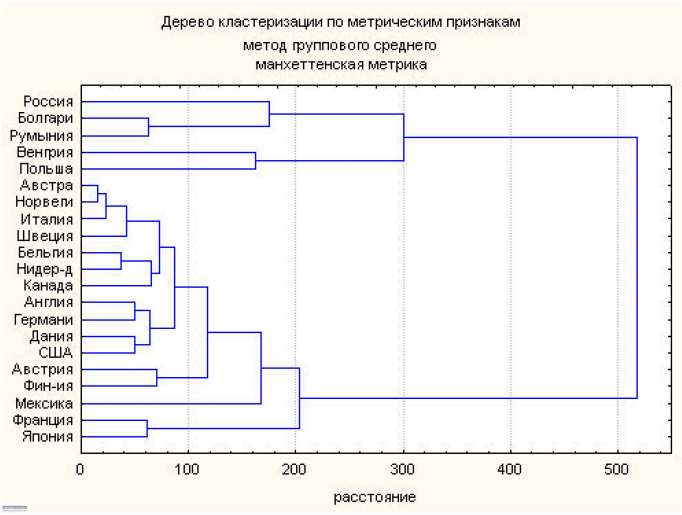


Рис. 18. Дерево кластеризации индивидов в случае исходных метрических признаков.

Для того чтобы все признаки участвовали в кластеризации индивидов, полезно исходные данные центрировать (вычитание математического ожидания) и нормировать (деление на стандартное отклонение). В этом

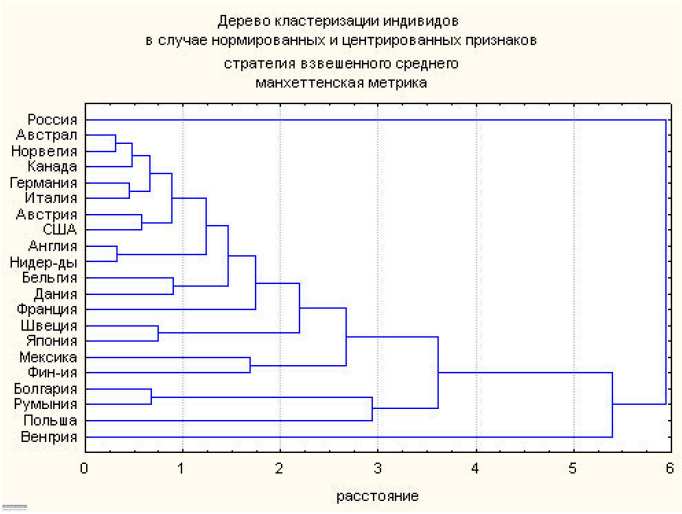


Рис. 19. Дерево кластеризации индивидов в случае нормированных и центрированных метрических признаков.

случае кластеризация осуществляется с учетом вклада всех признаков, а не только признака с наибольшим масштабом измерения. Выбор манхеттенской метрики и стратегии группового среднего для нормированных и центрированных данных приводит дереву кластеризации, в котором минимальным расстоянием между собой отличаются страны: Австралия-Норвегия и Англия-Нидерланды.

11.4. Кластеризация признаков

Разбиение признаков на кластеры можно осуществлять также как и в случае индивидов. Для этого нужно только транспонировать матрицу наблюдений. Наряду с перечисленными метриками используются также метрики, связанные с коэффициентом корреляции. Напомним, что коэффициент корреляции является характеристикой меры линейной зависимости между признаками ξ и η:

$$\varrho = \frac{E(\xi - E\xi)(\eta - E\eta)}{\sqrt{E(\xi - E\xi)^2}\sqrt{E(\eta - E\eta)^2}}.$$

Если признаки не зависимы, то коэффициент корреляции равен нулю. Если

признаки связаны линейной зависимостью вида $\eta = k\xi + b$, то коэффициент корреляции равен ± 1 в зависимости от знака коэффициента k.

При наличии матрицы наблюдений выборочный коэффициент кореляции между первым и вторым признаками вычисляется по формуле:

$$r = \frac{\sum\limits_{j=1}^{n} x_{1j}x_{2j} - n\bar{x_1}\bar{x_2}}{\sqrt{\sum\limits_{j=1}^{n} x_{1j}^2 - n\bar{x_1}^2}\sqrt{\sum\limits_{j=1}^{n} x_{2j}^2 - n\bar{x_2}^2}},$$

где также считаем, что имеется n индивидов и p признаков, x_{ij} – наблюдение i-го признака и j-го индивида, $\bar{x}_i$ – выборочное среднее i-го признака.

В качестве меры сходства между признаками выберем метрику, равную $1 - r^2$. Чем ближе коэффициент корреляции к нулю, тем больше метрика. Чем меньше метрика, тем больше связаны линейной зависимостью, т.е. можно сказать, тем меньше расстояние между признаками.

ПРИМЕР 1. Рассмотрим данные о жизни России за последние 30 лет.

Года	Продолж-ть жизни L	Кол-во (тыс.чел.) чиновников M	Водка (бут.в год) V	Доходы бедных P	Автомобили (шт.на 1000 чел.) A
1970	68.9	1060	25.3	7.8	5.5
1975	68.1	1101	28	9.5	15.3
1980	67.6	1147	30	10.1	30.2
1985	69.2	1204	23.5	10	44.5
1990	69.2	1602	18	9.8	58.6
1995	64.6	1893	38.4	5.5	93.3
1998	67	2777	29.6	6.2	122

Корреляционная матрица и матрица метрик вида $1 - r^2$ имеют вид:

r	L	M	V	P	A
L	1				
M	-0.5008	1			
V	-0.9341	0.2966	1		
P	0.7692	-0.7001	-0.6809	1	
A	-0.5953	0.9537	0.3701	-0.6692	1

$1-r^2$	L	M	V	P	A
L	0				
M	0.749	0			
V	0.128	0.912	0		
P	0.408	0.510	0.536	0	
A	0.646	**0.091**	0.863	0.552	0

Находим минимальное расстояние между признаками. Оно равно 0.091 и соответствует признакам: количество чиновников и количество автомобилей. Объединяем эти признаки в один, называемый чиновники-автомобили (AM) и таким образом получаем не пять, а четыре признака: L, V, P, AM. Расстояния между первыми тремя известны. Вычислим расстояния между AM и остальными признаками на основе гибкой стратегии, согласно которой имеем:

$$d_{AM,L} = 0.625(d_{A,L} + d_{M,L}) - 0.25d_{A,M} = 0.625(0.646 + 0.749) - 0.25 \cdot 0.091 = 0.849,$$

$$d_{AM,V} = 0.625(d_{A,V} + d_{M,V}) - 0.25d_{A,M} = 0.625(0.863 + 0.912) - 0.25 \cdot 0.091 = 1.087,$$

$$d_{AM,P} = 0.625(d_{A,P} + d_{M,P}) - 0.25d_{A,M} = 0.625(0.552 + 0.510) - 0.25 \cdot 0.091 = 0.641.$$

d	L	V	P	AM
L	0			
V	**0.128**	0		
P	0.408	0.510	0	
A	0.849	1.087	0.641	0

Находим минимальное расстояние между признаками, это расстояние между продолжительностью жизни L и количеством продаваемой водки V (0.128). Получаем три признака: доходы бедных P, водка-продолжительность жизни VL и автомобили-чиновники AM. Если задача состоит в нахождении трех кластеров признаков, то все получено. Если нужно определить, к какому из кластеров: водка-продолжительность жизни VL или автомобили-чиновники AM относится признак "доходы бедных"P, то продолжаем. Расстояние $d_{AM,P} = 0.641$ из вычислений на предыдущем шаге.

$$d_{VL,P} = 0.625(d_{V,P} + d_{L,P}) - 0.25d_{V,L} = 0.625(0.536 + 0.408) - 0.25 \cdot 0.128 = 0.559$$

$$d_{VL,AM} = 0.625(d_{V,AM} + d_{L,AM}) - 0.25d_{V,L} = 0.625(1.087 + 0.849) - 0.25 \cdot 0.128 = 1.178.$$

Отсюда получаем, что доходы бедных лучше прогнозируются по продолжительности жизни и продаваемой водке, нежели по количеству автомобилей и чиновников.

ПРИМЕР 2. Если в качестве метрики рассматривать характеристику $1-r$, показывающую степень удаленности от положительной корреляции, то результат получится несколько иной. Матрица расстояний имеет вид:

$1-r$	L	M	P	A	V
L	0				
M	1.50	0			
P	0.23	1.70	0		
A	1.60	**0.05**	1.67	0	
V	1.93	0.70	1.68	0.63	0

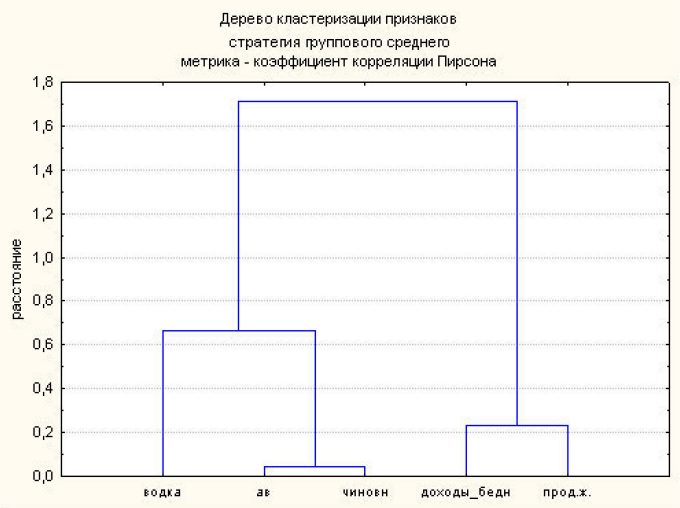


Рис. 20. Дерево кластеризации метрических признаков.

Также близкими являются признаки A - автомобили и M - чиновники.

d	AM	L	P	VM
AM	0			
L	1.92	0		
P	2.09	**0.23**	0	
V	0.82	1.93	1.68	0

$$d_{AM,L} = 0.625(d_{A,L} + d_{M,L}) - 0.25d_{A,M} = 0.625(1.6 + 1.5) - 0.25 \cdot 0.05 = 1.92,$$

$$d_{AM,V} = 0.625(d_{A,V} + d_{M,V}) - 0.25d_{A,M} = 0.625(0.63 + 0.7) - 0.25 \cdot 0.05 = 0.82,$$

$$d_{AM,P} = 0.625(d_{A,P} + d_{M,P}) - 0.25d_{A,M} = 0.625(1.67 + 1.7) - 0.25 \cdot 0.05 = 2.09.$$

Находим, что минимальное расстояние оказывается между продолжительностью жизни L и доходами бедных P.

Итак, имеется пара признаков AM и LP. К какой паре можно отнести количество продаваемой водки?

$$d_{V,LP} = 0.625(d_{V,L} + d_{V,P}) - 0.25d_{L,P} = 0.625(1.93 + 0.70) - 0.25 \cdot 0.23 = 2.2$$
$$d_{LP,AM} = 0.625(d_{L,AM} + d_{P,AM}) - 0.25d_{L,P} = 0.625(1.92 + 2.09) - 0.25 \cdot 0.23 = 2.45,$$
$$d_{V,AM} = 0.82.$$

Минимальное расстояние оказывается между признаками автомобили-чиновники и количеством продаваемой водки. Автоматически дерево кластеризации можно получить, используя стандартный пакет статистических программ StatSoft, установив в меню в качестве метрики коэффициент корреляции Пирсона, а в качестве стратегии, например, метод группового среднего.

12. Приложения

12.1. Теорема Фишера-Кочрена

Теорема 5. *(Fisher-Cochran) Пусть* $Y = (y_1, \ldots, y_n)$ *с независимыми компонетами,* $y_i \sim \mathcal{N}(0,1)$*, квадратичные формы* $Q_1, \ldots, Q_k$ *рангов* $n_1, \ldots, n_k$*, такие что*

$$Y^T Y = Q_1 + \ldots + Q_k\,.$$

Тогда $n = n_1 + \ldots + n_k$ *равносильно* $Q_i \sim \chi^{\mathbf{2}}(\mathbf{n_i})$ *и* $Q_1, \ldots, Q_k$ *независимы.*

Доказательство.

Пусть $Q_i = Y^T A_i Y$, где матрица A_i ранга n_i. Тогда существует матрица B_i размерности n_i на n, при помощи которой квадратичная форма приводится к диагональному виду, то есть $Q_i = (B_i Y)^T \Delta_i (B_i Y)$, где Δ_i диагональная матрица с элементами ± 1. Соберем k матриц B_i в одну матрицу B размерности n на n.

$$[B_1^T(n, n_1)|\ldots|B_k^T(n, n_k)] \cdot \begin{bmatrix} \Delta_1 & \ldots & \mathbf{0} \\ \vdots & \ldots & \vdots \\ \mathbf{0} & \ldots & \Delta_k \end{bmatrix} \cdot \begin{bmatrix} B_1(n_1, n) \\ \vdots \\ B_k(n_k, n) \end{bmatrix} = \sum_i B_i^T \Delta_i B_i$$

Следовательно,

$$Y^T Y = \sum_i Q_i = \sum_i (B_i Y)^T \Delta_i (B_i Y) =$$

$$= Y^T (\sum_i B_i^T \Delta_i B_i) Y = Y^T B^T \Delta B Y,\ \Rightarrow\ B^T \Delta B = \mathbf{I}\,.$$

Ранг матрицы B должен быть равен n, так как с одной стороны, он не может быть больше n, а с другой ранг произведения матриц не превышает минимального ранга сомножителей,

$$n = rank((B^T \Delta)B) \leq \min(rank(B^T \Delta), rank(B).$$

Тогда $\Delta = (B^T)^{-1}B^{-1}$ положительно определенная матрица, следовательно, $\Delta = \mathbf{I}$ и матрица B является ортогональной. Тогда компоненты вектора $X = BY$ независимы и нормальны, кроме того $Y^TY = X^TX$.

$$Q_1 = x_1^2 + \dots x_{n_1}^2,$$
$$Q_2 = x_{n_1+1}^2 + \dots x_{n_1+n_1}^2,$$
$$Q_3 = x_{n_1+n_2+1}^2 + \dots x_{n_1+n_2+n_3}^2,$$
$$\dots$$

Осюда все Q_i независимы и имеют распределение $\chi^{\mathbf{2}}(\mathbf{n_i})$ соответственно. Таким образом достаточность $n = n_1 + \dots + n_k$ установлена. Необходимость очевидна.

Следствие 1. *Для того чтобы квадратичная форма Y^TAY имела распределение χ^2, необходимо и достаточно условие идемпотентности матрицы $A = A^2$, причем $df = rank(A) = \mathrm{Tr}A$.*

Достаточность. Применим теорему Фишера-Кочрена к квадратичной форме ранга n

$$Y^TY = Y^TAY + Y^T(I - A)Y$$

и воспользуемся неравенством Сильвестра

$$rank(A) + rank(C) - rank(AC) \leq n$$

для идемпотентной матрицы $A = A^2$ и $C = \mathbf{I} - A$, то

$$rank(A) + rank(\mathbf{I} - A) - \underbrace{rank(A(\mathbf{I} - A))}_{=0} \leq n$$
$$n = rank(\mathbf{I}) = rank(A + \mathbf{I} - A) \leq rank(A) + rank(\mathbf{I} - A)\,,$$

Таким образом,

$$n = rank(A) + rank(\mathbf{I} - A),$$

и по теореме Фишера-Кочрена $Y^TAY \sim \chi^{\mathbf{2}}(\mathbf{rank}(\mathbf{A}))$.

Для доказательства необходимости заметим, что существует ортогональная матрица C такая, что при преобразовании $Y = CX$

$$Y^TAY = X^TC^TACX = \lambda_1 x_1^2 + \ldots + \lambda_m x_m^2,$$
$$Y^TY = X^TX = x_1^2 + \ldots + x_n^2,$$

где $\lambda_1, \ldots, \lambda_m$ ненулевые собственные числа матрицы A, $m = rank(A)$. Так как $Y^TAY \sim \chi^2(p)$, то характеристическая функция суммы $\lambda_1 x_1^2 + \ldots + \lambda_m x_m^2$ имеет вид

$$[(1 - 2i\lambda_1 t)\ldots(1 - 2i\lambda_m t)]^{-\frac{1}{2}} = (1 - 2it)^{-\frac{p}{2}}.$$

Сравнение х.ф. дает $p = m$, все $\lambda_i = 1$, C^TAC - диагональная матрица с нулями или единицами на главной диагонали.

$$C^TAC = C^TACC^TAC = C^TA^2C, \implies A^2 = A.$$

Число степеней свободы равно $p = m = rank(A)$, то есть $\mathrm{Tr}A$, так как $A^2 = A$.