Анализ корреляционного отношения

Точечная оценка показателя $r_{\eta\xi}$. Пусть *экспериментальные данные* представлены в форме (6.10), т.е. сгруппированы по значениям x_i случайной величины ξ.

Тогда за значение *точечной оценки* величины σ_f^2 принимают

$$\widehat{\sigma}_f^2 = \frac{1}{n}\sum_{i=1}^{m} n_i(\overline{y}_i - \overline{y})^2.$$

Значение точечной оценки дисперсии σ_η^2 находим по известной формуле (см. **2.1**)

$$\widehat{\sigma}_\eta^2 = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2.$$

Отсюда на основании (6.5) получаем значение точечной оценки показателя $r_{\eta\xi}$:

$$\widehat{r}_{\eta\xi} = \sqrt{\frac{\widehat{\sigma}_f^2}{\widehat{\sigma}_\eta^2}}. \tag{6.20}$$

Напомним, что точечная оценка $\widehat{r}(\vec{X}_n, \vec{Y}_n)$ определяет степень зависимости случайной величины η от случайной величины ξ. Аналогично можно ввести точечную оценку $\widehat{r}_{\xi\eta}$ для *корреляционного отношения* $r_{\xi\eta}$.

Пусть экспериментальные данные получены в форме (6.9) и не допускают удовлетворительной группировки по оси значений ξ (так как недостаточно велико n или точки (x_i, y_i) слишком „разрежены“ на плоскости).

В этом случае нужно выдвинуть некоторое предположение (*статистическую гипотезу*) о виде функции регрессии $\mathrm{M}(\eta|\xi = x) = f(x)$. Проверка таких гипотез будет рассмотрена ниже (см. **7**).

Допустим, что параметрический вид этой функции задан, т.е. принято предположение о том, что

$$f(x) = f(x;\theta_1,\ldots,\theta_k)$$

и найдены значения $\widehat{\theta}_i$ оценок параметров θ_i, $i = \overline{1,k}$ (см. 7). Тогда значение точечной оценки $\widehat{\sigma}_\eta^2$ для дисперсии σ_η^2 находим по формуле

$$\widehat{\sigma}_\eta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2,$$

а значение $\widehat{\overline{\sigma}}_\eta^2$ оценки $\overline{\sigma}_\eta^2$ можно записать в виде

$$\widehat{\overline{\sigma}}_\eta^2 = \frac{1}{n-k}\sum_{i=1}^{n}\Big(y_i - f\big(x_i;\widehat{\theta}_1,\ldots,\widehat{\theta}_k\big)\Big)^2. \tag{6.21}$$

Следовательно, согласно (6.5), точечную оценку показателя $r_{\eta\xi}$ можно определить равенством

$$\widehat{r}_{\eta\xi} = \sqrt{\frac{1-\widehat{\overline{\sigma}}_\eta^2}{\widehat{\sigma}_\eta^2}}. \tag{6.22}$$

Интервальная оценка и проверка значимости $r_{\eta\xi}$. Построение *доверительного интервала* для показателя $r_{\eta\xi}$ основано на том, что *статистика**

$$W = \frac{(n-m)\widehat{r}_{\eta\xi}^2(\vec{X}_m,\vec{Y}_n)}{(m-1)(1-\widehat{r}_{\eta\xi}^2(\vec{X}_m,\vec{Y}_n))}\,\frac{m-1}{m-1+nr_{\eta\xi}^2(\vec{X}_m,\vec{Y}_n)}$$

приближенно имеет *распределение Фишера* с числом степеней свободы r_1^* и $r_2 = n-m$, где

$$r_1^* = \frac{(m-1+n\widehat{r}_{\eta\xi}^2(\vec{X}_m,\vec{Y}_n))^2}{m-1+2n\widehat{r}_{\eta\xi}^2(\vec{X}_m,\vec{Y}_n)}, \tag{6.23}$$

в предположении, что при условии $\xi = x$ случайная величина η имеет нормальный закон распределения с постоянной дисперсией для любого x.

Используя квантили $F_{\alpha/2}(r_1^*,r_2)$ и $F_{1-\alpha/2}(r_1^*,r_2)$ распределе-н

ния Фишера для $\alpha = 1 - \gamma$, где γ — заданная *доверительная вероятность*, можно записать границы доверительного интервала в следующем виде:

$$\underline{r}_{\eta\xi} = \sqrt{\frac{(n-m)\hat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n)}{n(1-\hat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n))F_{\alpha/2}(r_1^*, r_2)} - \frac{m-1}{n}}, \quad (6.24)$$

$$\overline{r}_{\eta\xi} = \sqrt{\frac{(n-m)\hat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n)}{n(1-\hat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n))F_{1-\alpha/2}(r_1^*, r_2)} - \frac{m-1}{n}}. \quad (6.25)$$

Проверка значимости показателя $r_{\eta\xi}$ (т.е. проверка статистической гипотезы $H_0 : r_{\eta\xi} = 0$) основана на том*, что *статистика*

$$W_0 = \frac{(n-m)\hat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n)}{(m-1)(1-\hat{r}^2_{\eta\xi}(\vec{X}_m, \vec{Y}_n))} \quad (6.26)$$

имеет распределение Фишера с числом степеней $r_1 = m - 1$ и $r_2 = n - m$, если гипотеза $H_0: r_{\eta\xi} = 0$ верна.

Границу *критического множества* для гипотезы $H_0: r_{\eta\xi} = 0$ на *уровне значимости* α определяет квантиль $f_{1-\alpha}(r_1, r_2)$. Величину показателя $r_{\eta\xi}$ следует считать значимо отличающейся от нуля, если значение статистики W_0 принадлежит критическому множеству, т.е. ее значение больше $f_{1-\alpha}(r_1, r_2)$. В противном случае делаем вывод об отсутствии *стохастической связи* между η и ξ.

Пример 6. 7. Пусть в результате обработки $n = 132$ экс-периментальных точек (x_i, y_i), $i = 1, n$, получено *выборочное значение* корреляционного отношения $\hat{r}_{\eta\xi} = 0{,}6\ 0$, причем про-межуток, содержащий все *выборочные значения* случайной ве-личины ξ, был разбит на $m = 12$ равных интервалов (см. **1.3**). Найдем значения границ доверительного интервала $(\underline{r}, \overline{r})$ для показателя $r_{\eta\xi}$ с уровнем доверия $\gamma = 0{,}9$ и проверим значи-мость этого показателя на уровне значимости $\alpha = 0{,}1$.

Сначала определим по формуле (6.23) число степеней свободы r_1^* (округляя до целого числа):

$$r_1^* = \frac{(12 - 1 + 132 \cdot 0{,}36)^2}{12 - 1 + 2 \cdot 132 \cdot 0{,}36} \approx 27.$$

По таблице квантилей распределения Фишера с числом степеней свободы $r_1^* = 27$ и $r_2 = n - m = 132 - 12 = 120$ (см. табл. П.4) находим квантили уровней $\alpha/2 = (1-\gamma)/2 = 0{,}05$ и $1 - \alpha/2 = 0{,}95$:

$$f_{0,05}(27, 120) = 1{,}58;$$

$$f_{0,95}(27, 120) = \frac{1}{f_{0,05}(120, 27)} = \frac{1}{1{,}73} \approx 0{,}58.$$

По формулам (6.24), (6.25) находим значения границ доверительного интервала:

$$\underline{r} = \sqrt{\frac{120 \cdot 0{,}36}{132 \cdot 0{,}64 \cdot 1{,}58} - \frac{11}{132}} = 0{,}49,$$

$$\overline{r} = \sqrt{\frac{120 \cdot 0{,}36}{132 \cdot 0{,}64 \cdot 0{,}58} - \frac{11}{132}} = 0{,}93.$$

Таким образом, с вероятностью $\gamma = 0{,}9$ истинное значение показателя $r_{\eta\xi}$ (при точечной оценке $\widehat{r}_{\eta\xi} = 0{,}60$) заключено в пределах $0{,}49 < r_{\eta\xi} < 0{,}93$.

Для проверки значимости $r_{\eta\xi}$ (хотя она и так очевидна) найдем квантиль распределения Фишера $f_{1-\alpha}(r_1, r_2)$ при $\alpha = 0{,}1$, $r_1 = 11$, $r_2 = 120$. Поскольку $f_{0,9}(120, 11) = 1{,}58$, то $f_{0,1}(11, 120) = 1/f_{0,9}(120, 11) = 0{,}63$. Значение статистики W_0 равно $6{,}1 > f_{0,1} = 0{,}63$, следовательно, гипотеза H_0: $r_{\eta\xi} = 0$ уверенно отклоняется, т.е. между переменными ξ и η имеет место стохастическая связь.

Множественный коэффициент корреляции. Для того чтобы результаты, были частным случаем рассматриваемой общей ситуации, сохраним обозначение η для

„выходной“ переменной X_0 и обозначение ξ для „входной“ переменной, но теперь ξ будет вектором размерности p, т.е. $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)$. Возможные значения переменной η будем обозначать y, а возможные значения $\vec{\xi}$ — $\vec{x} = (x_1, \ldots, x_p)$.

При решении практических задач, связанных с анализом стохастических связей между многими случайными переменными, чаще других рассматривают ситуацию, в которой поведение какой-то одной (выходной) переменной η стараются объяснить поведением совокупности других (входных) переменных $\vec{\xi} = (\xi_1, \ldots, \xi_p)$.

Прежде всего убедимся, что наилучшим прогнозом (аппроксимацией) для неизвестного значения η (в смысле средней квадратичной ошибки) является условное математическое ожидание η при условии $\vec{\xi} = \vec{x}$, т.е. величина $\mathrm{M}(\eta | \vec{\xi}) = \vec{x}) = f(\vec{x})$, где $\vec{x} = (x_1, \ldots, x_n)$.

Действительно, пусть $\tilde{f}(\vec{x})$ — любая функция. Тогда

$$\mathrm{M}(\eta - \tilde{f}(\vec{\xi}))^2 = \mathrm{M}((\eta - f(\vec{\xi})) + (f(\vec{\xi}) - \tilde{f}(\vec{\xi})))^2 =$$

$$= \mathrm{M}(\eta - f(\vec{\xi}))^2 + \mathrm{M}(f(\vec{\xi}) - \tilde{f}(\vec{\xi}))^2 + 2\mathrm{M}((f(\vec{\xi}) - \tilde{f}(\vec{\xi}))(\eta - f(\vec{\xi}))).$$

Поскольку последнее слагаемое равно нулю (доказательство этого аналогично тому, которое приведено в **6.2**), то

$$\min \mathrm{M}(\eta - \tilde{f}(\vec{\xi}))^2 = \mathrm{M}(\eta - f(\vec{\xi}))^2,$$

если $\tilde{f}(\vec{\xi}) = f(\vec{\xi})$. Следовательно, при каждом данном значении $\vec{\xi} = \vec{x}$ и любой функции $\tilde{f}(\vec{x}) \neq f(\vec{x})$ имеет место неравенство

$$\mathrm{M}(\eta - \tilde{f}(\vec{x}))^2 > \mathrm{M}(\eta - f(\vec{x}))^2.$$

Таким образом, мы снова (как и в **6.1**) пришли к *функции регрессии* $f(\vec{x}) = \mathrm{M}(\eta | \vec{\xi} = \vec{x})$, но уже функции от p переменных $x_1, \ldots, x_p$, которая наиболее точно (в смысле сред-

ней квадратичной ошибки) воспроизводит значения исследуемого результирующего переменного η по заданным величинам $\vec{x} = (x_1, \ldots, x_p)$ входных переменных $\vec{\xi} = (\xi_1, \ldots, \xi_p)$.

Теперь вернемся к соотношению (6.4), которое связывает дисперсию σ_η^2 случайной величины η с величинами $\sigma_f^2 = \mathbf{D} f(\xi)$ и $\overline{\sigma}_\eta^2 = \mathbf{M}\mathbf{D}(\eta|\xi)$. Соотношение (6.4) остается справедливым и в случае вектора входных переменных $\vec{\xi} = (\xi_1, \ldots, \xi_p)$.

Следовательно, так же как и в случае парной зависимости, случайный разброс (вариация) выходного переменного η складывается из контролируемой нами (посредством $\vec{x} = (x_1, \ldots, x_p)$) вариации функции регрессии $f(\vec{x})$ и из неподдающегося нашему контролю случайного разброса значений η (при фиксированном $\vec{x}$) относительно функции регрессии. Именно этот неконтролируемый разброс определяет меру зависимости переменной η от переменной $\vec{\xi}$, которая характеризуется величиной $\overline{\sigma}_\eta^2$. Чем меньше значение $\overline{\sigma}_\eta^2$, тем точнее прогноз. При $\overline{\sigma}_\eta^2 = 0$ случайные величины η и $\vec{\xi}$ связаны функциональной зависимостью.

Эти соображения подводят нас к определению ***множественного коэффициента корреляции*** R_η, под которым понимают величину

$$R_\eta = \sqrt{1 - \frac{\overline{\sigma}_\eta^2}{\sigma_\eta^2}}. \tag{6.31}$$

Заметим, что квадрат R_η^2 показателя R_η принято называть ***коэффициентом детерминации.***

Покажем, что R_η есть коэффициент корреляции между η и $f(\vec{\xi})$ (тем самым оправдаем его название). Имеем

$$\begin{aligned}
\mathbf{cov}(\eta, f(\vec{\xi})) &= \mathbf{M}\big((\eta - \mathbf{M}\eta)(f(\vec{\xi}) - \mathbf{M}\eta)\big) = \\
&= \mathbf{M}\big((f(\vec{\xi}) - \mathbf{M}\eta)^2 + (\eta - f(\vec{\xi}))(f(\vec{\xi}) - \mathbf{M}\eta)\big) = \\
&= \mathbf{M}\big(f(\vec{\xi}) - \mathbf{M}\eta\big)^2 + \mathbf{M}\big((\eta - f(\vec{\xi}))(f(\vec{\xi}) - \mathbf{M}\eta)\big) = \sigma_f^2,
\end{aligned}$$

поскольку

$$\mathbf{M}\big((\eta - f(\vec{\xi}))(f(\vec{\xi}) - \mathbf{M}\eta)\big) = 0.$$

Далее,

$$R_\eta = \sqrt{1 - \frac{\overline{\sigma}_\eta^2}{\sigma_\eta^2}} = \sqrt{\frac{\sigma_\eta^2 - \overline{\sigma}_\eta^2}{\sigma_\eta^2}} = \sqrt{\frac{\sigma_f^2}{\sigma_\eta^2}} = \frac{\sigma_f^2}{\sqrt{\sigma_\eta^2 \sigma_f^2}} = \frac{\mathbf{cov}(\eta, f(\vec{\xi}))}{\sigma_\eta \sigma_f}.$$

Отметим свойства показателя R_η, которые непосредственно вытекают из соотношения (6.31), справедливого и в многомерном случае.

1°. $0 \leqslant R_\eta \leqslant 1$.

2°. $R_\eta = 0$ соответствует $\sigma_f^2 = \mathbf{D} f(\vec{\xi}) = 0$. В частности, функция регрессии f не зависит от значений ее аргументов $\vec{x}$: $f(\vec{x}) = \text{const}$.

3°. $R_\eta = 1$ соответствует $\overline{\sigma}_\eta^2 = 0$ и означает наличие чисто функциональной связи между η и $\vec{\xi} = (\xi_1, \ldots, \xi_p)$: $\eta = f(\xi_1, \ldots, \xi_p)$.

Определение показателя R_η в виде (6.31) и отмеченные свойства 1°–3° справедливы при любом законе распределения вектора $(\eta, \xi_1, \ldots, \xi_p)$.

Если же предположить, что исходные *статистические данные* $(x_{1i}, x_{2i}, \ldots, x_{pi})$, y_i, $i = \overline{1, n}$, могут интерпретироваться как выборка объема n из $(p+1)$-мерной *генеральной совокупности*, распределенной по нормальному закону с вектором средних значений $\vec{\mu} = (\mu_0, \mu_1, \ldots, \mu_p)$, где $\mu_0 = \mathbf{M}\eta$, $\mu_i = \mathbf{M}\xi_i$, $i = \overline{1, p}$, и ковариационной матрицей Σ, то можно отметить дополнительные свойства показателя R_η и правила его вычисления.

Прежде всего укажем на то, что в рассматриваемой ситуации (ср. с примером 6.4) условное математическое ожидание η при фиксированных значениях $\xi_1 = x_1, \ldots, \xi_p = x_p$ (т.е. функция регрессии $f(x)$) является линейной функцией переменных

$x_1, \ldots, x_p$, а условная дисперсия $\mathbf{D}(\eta|\vec{\xi}=\vec{x})$ не зависит от $\vec{x}=(x_1,\ldots,x_p)$ и имеет вид

$$\mathbf{D}(\eta)\vec{\xi}=\vec{x}=\sigma_\eta^2(1-R_\eta^2).$$

Последнее выражение — полная аналогия формулы (6.2), только роль коэффициента корреляции ρ играет множественный коэффициент корреляции R_η.

Приведем без доказательства* следующие дополнительные свойства показателя R_η в случае совместного нормального закона распределения переменных η и $\vec{\xi}=(\xi_1, \ldots, \xi_p)$.

4°. С помощью корреляционной матрицы P (6.27) показатель R_η можно вычислить по формуле

$$R_\eta=\sqrt{1-\frac{\det P}{P_{00}}}, \tag{6.32}$$

где $\det P$ — определитель матрицы P, а P_{00} — алгебраическое дополнение элемента $\rho_{00}=1$.

5°. Показатель R_η можно вычислить, используя частные коэффициенты корреляции следующим образом:

$$R_\eta^2=1-(1-\rho_{01}^2)\prod_{j=2}^{p}(1-\rho_{0j(12\ldots j-1)}^2). \tag{6.33}$$

6°. Множественный коэффициент корреляции мажорирует любой парный коэффициент корреляции, характеризующий стохастическую связь результирующего показателя η с остальными, т.е.

$$|\rho_{0j}|\leqslant R_\eta, \quad |\rho_{0j(\cdot)}\leqslant R_\eta, \quad j=\overline{0,p},$$

где $\rho_{0j(\cdot)}$ — произвольный частный коэффициент корреляции, содержащий нуль среди первичных индексов.

7°. Присоединение каждого нового предсказывающего (входного) переменного не может уменьшить величины R_η (независимо от порядка присоединения).

Статистический анализ множественного коэффициента корреляции. Вычисление значений точечной оценки $\widehat{R}_\eta$ показателя R_η проводится по тем же формулам (6.31)–(6.33) путем подстановки в них вместо значений *теоретических характеристик* соответствующих значений *выборочных характеристик*.

Например, при использовании формулы (6.32) матрицу P нужно заменить матрицей $\widehat{P}$, в которой все элементы ρ_{ij} заменены на $\widehat{\rho}_{ij}$, $i,j=\overline{0,p}$, а при использовании формулы (6.33) коэффициент корреляции ρ_{01} и все частные коэффициенты корреляции $\rho_{ij(\cdot)}$ нужно заменить значениями $\widehat{\rho}_{ij(\cdot)}$.

Для проверки гипотезы H_0: $R_\eta=0$ будем предполагать, что случайный вектор (ξ,η) имеет $(p+1)$-мерный нормальный закон распределения, и воспользуемся тем*, что *статистика*

$$W_1=\frac{\widehat{R}_\eta^2}{1-\widehat{R}_\eta^2}\,\frac{n-p-1}{p}$$

имеет *распределение Фишера* с p и $n-p-1$ степенями свободы, если истинное значение $R_\eta=0$.

Гипотеза об отсутствии множественной корреляционной связи между η и $\vec{\xi}=(xi_1,\ \ldots,\ \xi_p)$ отвергается на уровне значимости α, если

$$\frac{\widehat{R}^2}{1-\widehat{R}^2}\,\frac{n-p-1}{p}>F_{1-\alpha}(p,n-p-1). \tag{6.34}$$

В предположении, что η при условии $\vec{\xi}=\vec{x}$ имеет нормальный закон с постоянной дисперсией для любого $\vec{x}$, можно показать**, что значения приближенных доверительных границ $\underline{R}_\eta$

и $\underline{R}_\eta$ для показателя R_η, отвечающие доверительной вероятности $\gamma = 1 - \alpha$ и выборке объема n, имеют вид (справедливый при условии $p \geqslant 8$):

$$\underline{R}_\eta = \sqrt{\frac{\widehat{R}_\eta^2[1-(p+1)/n]}{(1-\widehat{R}_\eta^2)F_{1-\alpha/2}(r_1, r_2)} - \frac{p}{n}}, \tag{6.35}$$

$$\overline{R}_\eta = \sqrt{\frac{\widehat{R}_\eta^2[1-(p+1)/n]}{(1-\widehat{R}_\eta^2)F_{\alpha/2}(r_1, r_2)} - \frac{p}{n}}, \tag{6.36}$$

где

$$r_1 = \frac{(p+n\widehat{R}_\eta)^2}{p+2n\widehat{R}_\eta^2}, \quad r_2 = n-p-1.$$

Пример 6.10. Вернемся к примерам 6.8 и 6.9.

В примере 6.8 найдем значения оценок множественного коэффициента корреляции R_η между показателем качества η пряжи и совокупностью двух факторов: количеством ξ_1 профилактических наладок и числом ξ_2 обрывов нити.

Используя формулу (6.33), в которой вместо истинных значений показателей корреляции использованы значения их выборочных оценок (см. пример 6.3), получаем

$$\widehat{R}^2 = 1 - (1-\widehat{\rho}_{01}^2)(1-\widehat{\rho}_{02(1)}^2) =$$

$$= 1 - (1-0{,}105^2)(1-0{,}906^2) = 0{,}823,$$

откуда $\widehat{R} = \sqrt{0{,}823} = 0{,}907$.

В примере 6.9 найдем значения оценок показателя R_η множественной корреляции между урожайностью η кормовых трав и природными факторами: весенним количеством ξ_1 осадков и накопленной суммой ξ_2 температур.

Используя найденные в примере 6.4 оценки $\widehat{\rho}_{01} = 0{,}8$ и $\widehat{\rho}_{02(1)} = 0{,}097$, по той же формуле (6.33) находим (с заменой истинных значений показателей корреляции значениями их оценок.