1. Регрессия

Пусть случайные величины ξ и η имеют совместное распределение непрерывного типа с плотностью $f(x,y)$, для любого x частная плотность

$$f_1(x) = \int\limits_{-\infty}^{+\infty} f(x,y)dy > 0\,,$$

а условная плотность вероятности

$$f(y|x) = \frac{f(x,y)}{f_1(x)}$$

есть распределение массы в бесконечно узкой вертикальной полосе, проходящей через точку $\xi = x$. *Кривую регрессии для среднего величины* η определяют как траекторию точки $(x, \mathbf{E}(\eta|\xi = x)$ при переменном x:

$$y = \mathbf{E}(\eta|\xi = x)\,. \tag{1}$$

Кривые регрессии обладают свойством минимальности. Среди всех возможных функций $g(\xi)$ минимум выражения

$$\mathbf{E}(\eta - g(\xi))^2 = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} (y - g(x))^2 f(x,y)dxdy = \tag{2}$$
$$= \int\limits_{-\infty}^{+\infty} f_1(x)dx \int\limits_{-\infty}^{+\infty} (y - g(x))^2 f(y|x)dy$$

достигается при функции $g(\xi) = \mathbf{E}(\eta|\xi)$, изображаемой в виде кривой (1). Если кривая регрессии есть прямая, то имеет место случай *линейной регрессии.*

Если в экстремальной задаче (2) ограничиться функциями, принадлежащему определенному классу, например, линейными функциями, полиномами и т.д., то кривые, полученные этим путем, называются *кривыми средней квадратической регрессии.*

1.1. Оценка параметров линейной средней квадратической регрессии

Рассмотрим двумерную выборку наблюдений (x_i, y_i) случайных величин ξ и η, $i = 1, \ldots, n$. Из всего семейства прямых вида $g(x) = \alpha + \beta x$ выбирается „наилучшая“ в том смысле, что квадратичная форма

$$Q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

принимает наименьшее значение. В регрессионной модели

$$y_i = \alpha + \beta x_i + \delta_i \tag{3}$$

считаем x_i детерминированными, а δ_i – центрированными, некоррелированными случайными величинами с дисперсией σ^2,

$$\mathbf{E}\delta_i = 0, \quad \mathbf{D}\delta_i = \sigma^2, \quad \mathbf{E}\delta_i\delta_j = 0. \tag{4}$$

Случайные величины δ_i называют ошибками, так как они показывают, насколько велико отклонение наблюдений y_i от линии регрессии $y = \alpha + \beta x$.

Для оценки параметров α и β следует решить систему нормальных уравнений, полученную в результате дифференцирования квадратичной формы Q по параметрам и приравнивания полученных выражений нулю:

$$\begin{cases} -\frac{1}{2}\frac{\partial Q}{\partial \alpha} = \sum\limits_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \\ -\frac{1}{2}\frac{\partial Q}{\partial \beta} = \sum\limits_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0 \end{cases} \iff$$

$$\iff \begin{cases} \sum\limits_{i=1}^{n} y_i - \alpha n - \beta \sum\limits_{i=1}^{n} x_i = 0 \\ \sum\limits_{i=1}^{n} y_i x_i - \alpha \sum\limits_{i=1}^{n} x_i - \beta \sum\limits_{i=1}^{n} x_i^2 = 0 \end{cases}$$

Учитывая то, что $\sum\limits_{i=1}^{n} x_i = n\bar{x}$ и $\sum\limits_{i=1}^{n} y_i = n\bar{y}$, получаем оценки параметров линейной регрессии:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{5}$$

Не умаляя общности, будем считать

$$\sum_{i=1}^{n} x_i = 0. \tag{6}$$

При этом оценки параметров линейной регрессии упрощаются и имеют вид

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2}. \tag{7}$$

1.2. Распределение $\hat{\alpha}$ и $\hat{\beta}$ при известной дисперсии σ^2

Предложение 1. *Пусть ошибки регрессии распределены нормально, то есть* $\delta_i \sim \mathcal{N}(0;\sigma)$, $\sum\limits_{i=1}^{n} x_i = 0$. *Обозначим через* $[x,x]$ *скалярное произведение* $[x,x] = \sum\limits_{i=1}^{n} x_i^2$. *Тогда для оценок (7) справедливо*

$$1)\ \ \hat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma}{\sqrt{n}}\right), \qquad 2)\ \ \hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma}{\sqrt{[x,x]}}\right).$$

Действительно,

$$\hat{\alpha} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i + \delta_i) = \\ = \alpha + \frac{\beta}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\delta_i.$$

Так как $\sum\limits_{i=1}^{n} x_i = 0$ в силу (6), то получаем

$$\hat{\alpha} = \alpha + \bar{\delta}, \tag{8}$$

где $\bar{\delta} = \frac{1}{n}\sum\limits_{i=1}^{n}\delta_i$. Согласно (4), получаем

$$\mathbf{E}\hat{\alpha} = \alpha, \quad \mathbf{D}\hat{\alpha} = \sigma^2/n, \quad \hat{\alpha} \sim \mathcal{N}(\alpha, \sigma/\sqrt{n}).$$

Далее, выразим оценку $\hat{\beta}$ через ошибки δ_i. Подставим (3) в выражение (4) оценки $\hat{\beta}$.

$$\hat{\beta} = \frac{\sum\limits_i y_i x_i}{[x,x]} = \frac{\sum\limits_i (\alpha + \beta x_i + \delta_i)x_i}{[x,x]} = \\ = \frac{\alpha\sum\limits_i x_i + \beta\sum\limits_i x_i^2 + \sum\limits_i \delta_i x_i}{[x,x]}.$$

Из $\sum\limits_{i=1}^{n} x_i = 0$ и $[x,x] = \sum\limits_{i=1}^{n} x_i^2$ получаем

$$\hat{\beta} = \beta + \frac{1}{[x,x]}\sum_i \delta_i x_i. \tag{9}$$

Отсюда получаем выражения для моментов:

$$\mathbf{E}\hat{\beta} = \beta,$$

$$\mathbf{D}\hat{\beta} = \mathbf{D}\left(\beta + \frac{1}{[x,x]}\sum_i \delta_i x_i\right) = \frac{1}{[x,x]^2}\sum_i x_i^2 \mathbf{D}\delta_i = \frac{1}{[x,x]}\sigma^2.$$

Заметим, что

$$\mathbf{E}(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) = \mathbf{E}\left(\frac{1}{n}\sum_i \delta_i \cdot \frac{1}{[x,x]}\sum_j x_j\delta_j\right) = \\ = \frac{1}{n[x,x]}\sum_i x_i \mathbf{E}\delta_i^2 = \frac{\sigma^2}{n[x,x]}\sum_i x_i = 0\,.$$

1.3. Распределение $\hat{\alpha}$ и $\hat{\beta}$ в случае неизвестной дисперсии ошибок σ^2

Обозначим через $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ оценку линии регрессии и через

$$Q_E = \sum_i (y_i - \hat{y}_i)^2 \tag{10}$$

сумму квадратов отклонений наблюдаемых значаний y_i от линии регрессии $\hat{y}_i$. В качестве оценки дисперсии σ^2 используется выражение

$$\hat{\sigma}^2 = \frac{Q_E}{n-2} = S^2. \tag{11}$$

Предложение 2. *1) Отношение* $\frac{Q_E}{\sigma^2}$ *распределено по* $\chi^2(\mathbf{n-2})$. *2) Оценка* $\hat\sigma^2 = S^2$ *несмещенная,* $\mathbf{E}Q_E = (n-2)\sigma^2$.

Обозначим через $d_i = y_i - \hat y_i = y_i - \hat\alpha - \hat\beta x_i$. Используя то, что $\sum_i x_i = 0$, $\hat\alpha = \bar y$, $\hat\beta = \frac{\sum\limits_{i=1}^n x_i y_i}{\sum\limits_{i=1}^n x_i^2}$, убеждаемся, что

$$\sum_i d_i = \sum_i (y_i - \hat\alpha - \hat\beta x_i) = 0,$$
$$\sum_i x_i d_i = \sum_i x_i (y_i - \hat\alpha - \hat\beta x_i) = 0\,.$$

На основании этого

$$\sum_i \delta_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2 =$$
$$= \sum_i ((y_i - \hat\alpha - \hat\beta x_i) + (\hat\alpha - \alpha) + (\hat\beta - \beta)x_i)^2 =$$
$$= Q_E + n(\hat\alpha - \alpha)^2 + (\hat\beta - \beta)^2 [x,x].$$

При этом перекрестные суммы оказываются равными нулю:

$$\sum_i d_i(\hat\alpha - \alpha) = 0,$$
$$\sum_i d_i x_i (\hat\beta - \beta) = 0,$$
$$\sum_i (\hat\alpha - \alpha)(\hat\beta - \beta) x_i = 0\,.$$

Согласно (8), (9), лемме,[1] в которой в

$$x_i = \delta_i,$$
$$y_1 = \sqrt{n}(\hat\alpha - \alpha) = \sqrt{n}\bar\delta,$$
$$y_2 = \sqrt{[x,x]}(\hat\beta - \beta) = \frac{1}{\sqrt{[x,x]}} \sum_i \delta_i x_i,$$

квадратичная форма Q_E имеет ранг $n-2$, и $\frac{Q_E}{\sigma^2}$ имеет распределение хи-квадрат с $n-2$ степенями свободы.

---

[1]

Лемма 1. *(Фишера) Пусть* $x_1, \ldots, x_n$ *— независимые и нормально распределенные* $\mathcal{N}(0,\sigma)$ *компоненты вектора* X. *В результате ортогонального преобразования* $Y = CX$ *имеем также независимые и нормально распределенные* $\mathcal{N}(0,\sigma)$ *компоненты* $y_1, \ldots, y_n$. *Тогда квадратичная форма*

$$Q(X) = \sum_{i=1}^n x_i^2 - y_1^2 - \ldots - y_p^2$$

независима от $y_1, \ldots, y_p$ *и имеет плотность* $\frac{1}{\sigma^2}\kappa_{n-p}\left(\frac{x}{\sigma^2}\right)$, *где* $\kappa_n(x)$ *— плотность распределения хи-квадрат с* n *степенями свободы.*

Предложение 3. *Пусть* $\delta_i \sim \mathcal{N}(0;\sigma)$, *где дисперсия* σ^2 *неизвестна,*

$$S_\alpha^2 = S^2/n, \quad S_\beta^2 = S^2/[x,x],$$

$[x,x] = \sum\limits_{i=1}^{n} x_i^2$. *Тогда*

$$T' = \frac{\hat\alpha - \alpha}{S_\alpha} \sim \mathbf{T}(n-2), \tag{12}$$

$$T'' = \frac{\hat\beta - \beta}{S_\beta} \sim \mathbf{T}(n-2). \tag{13}$$

В случае нецентрированных данных $\sum\limits_{i=1}^{n} x_i \neq 0$ *вводятся изменения в*

$$[x,x] = \sum_{i=1}^{n}(x_i - \bar x)^2 \quad u$$

$$S_\alpha^2 = \frac{S^2 \sum\limits_{i=1}^{n} x_i^2}{n[x,x]}. \tag{14}$$

Доказательство. 1) Из предл. 1 имеем

$$U_1 = \frac{\hat\alpha - \alpha}{\sigma/\sqrt{n}} \sim \mathbf{N}(0;1).$$

Из предл. 2 и выражения (11) получаем, что

$$\chi^2 = \frac{(n-2)S^2}{\sigma^2} \sim \chi^{\mathbf{2}}(n-2).$$

Исходя из этого построим случайную величину T', имеющую распределение Стьюдента:

$$T' = \frac{U_1}{\sqrt{\frac{1}{n-2}\chi^2}} = \frac{(\hat\alpha - \alpha)\sqrt{n}}{\sigma \cdot \sqrt{\frac{1}{n-2}(n-2)S^2/\sigma^2}} =$$

$$= \frac{\hat\alpha - \alpha}{S}\sqrt{n} \sim \mathbf{T}(n-2).$$

2) Согласно предл. 1 имеем

$$U_2 = \frac{\hat\beta - \beta}{\sigma/\sqrt{[x,x]}} \sim \mathcal{N}(0;1).$$

Аналогично строим случайную величину

$$T'' = \frac{(\hat\beta - \beta)\sqrt{[x,x]}}{\sigma \cdot \sqrt{\frac{1}{n-2}(n-2)S^2/\sigma^2}} =$$

$$= \frac{\hat\beta - \beta}{S}\sqrt{[x,x]} \sim \mathbf{T}(n-2).$$

В случае нецентрированных данных $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. Дисперсия оценки параметра α имеет вид[2] $\mathbf{D}\hat{\alpha} = \mathbf{D}(\bar{y} - \hat{\beta}\bar{x}) = \mathbf{D}\bar{y} + \mathbf{D}(\hat{\beta}\bar{x}) =$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{[x,x]} = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} =$$

$$\frac{\sigma^2\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2\sigma^2}{n\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2\sum x_i^2}{n\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}.$$

Таким образом, по выборке дисперсии коэффициентов регрессии могут быть оценены следующим образом:

$$S_\alpha^2 = \frac{S^2\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}, \qquad S_\beta^2 = \frac{S^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}. \tag{15}$$

Выражения S_α и S_β называются стандартными ошибками коэффициентов регрессии.

1.4. Коэффициенты корреляции и детерминации

Для измерения доли информации, объясняемой регрессией, вводится *коэффициент детерминации*

$$R^2 = 1 - \frac{Q_E}{Q} = \frac{Q_R}{Q}, \tag{16}$$

где $Q = \sum\limits_{i=1}^{n}(y_i - \bar{y})^2$ – общая сумма квадратов отклонений от выборочного среднего.

---

[2] Здесь используется факт о том, что $\mathbf{cov}(\bar{y}, \hat{\beta}) = 0$. Это можно показать непосредственно, $\bar{y} = \alpha + \beta\bar{x} + \bar{\delta}$, $\mathbf{E}\bar{y} = \alpha + \beta\bar{x}$, $\bar{y} - \mathbf{E}\bar{y} = \bar{\delta}$. С другой стороны, $\hat{\beta} = \sum_i c_i(y_i - \bar{y}) = \sum_i c_i(\beta(x_i - \bar{x}) + (\delta_i - \bar{\delta}))$, где $c_i = \frac{x_i - \bar{x}}{[x,x]}$. $\mathbf{cov}(\bar{y}, \hat{\beta}) = \mathbf{E}\bar{\delta}(\hat{\beta} - \beta) = \mathbf{E}\bar{\delta}\hat{\beta} = \mathbf{E}\bar{\delta}\sum_i c_i(\beta(x_i - \bar{x}) + (\delta_i - \bar{\delta})) = 0$, так как $\mathbf{E}\bar{\delta}\delta_i = \frac{\sigma^2}{n}$, $\mathbf{E}\bar{\delta}^2 = \frac{\sigma^2}{n}$.

Предложение 4. *Обозначим через Q_R сумму квадратов отклонений линии регрессии от выборочного среднего:*

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

Тогда справедливо тождество

$$Q = Q_R + Q_E \tag{17}$$

Действительно, рассмотрим сумму квадратов

$$\begin{aligned} Q = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \\ = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \\ = Q_E + Q_R. \end{aligned}$$

Дополнительное слагаемое равно нулю. Это нетрудно показать. Так как $\hat{y}_i - \bar{y} = \hat{\beta}(x_i - \bar{x})$, то

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} \hat{\beta}(y_i - \hat{y}_i)(x_i - \bar{x}) = \\ = \hat{\beta} \sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))(x_i - \bar{x}) = \\ = \hat{\beta}(\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})^2) = 0, \end{aligned}$$

поскольку для оценки параметра $\hat{\beta}$ справедливо выражение

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{18}$$

Выборочный коэффициент корреляции, полученный по двумерной выборке $(x_1, y_1), \ldots, (x_n, y_n)$, вычисляется по формуле:

$$\hat{\varrho} = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2}} \tag{19}$$

Предложение 5. *Коэффициент детерминации равен квадрату коэффициента корреляции, $R^2 = \hat{\varrho}^2$.*

Действительно,

$$R^2 = 1 - \frac{Q_E}{Q} = \frac{Q_R}{Q} = \frac{\sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}^2 \sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}.$$

Подставляя (18) вместо $\hat{\beta}$ и сокращая на $\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$, получаем выражение для $\hat{\varrho}^2$.

$$R^2 = \frac{\left(\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum\limits_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left(\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 \sum\limits_{i=1}^{n}(y_i - \bar{y})^2} = \hat{\varrho}^2 .$$

1.5. Проверка гипотез в корреляционном анализе

Проверка значимости коэффициента детерминации, то есть гипотезы о том, что $H_0 : R^2 = 0$, эквивалентна проверке гипотез о равенстве нулю коэффициента корреляции $H_0 : \varrho = 0$ и коэффициента регрессии $H_0 : \beta = 0$. Все эти три гипотезы проверяются при помощи статистики Фишера вида

$$F = \frac{Q_R}{Q_E/(n-2)}, \tag{20}$$

которая в случае справедливости нулевых гипотез имеет распределение Фишера с 1 и $n-2$ степенями свободы. Из предл.2 следует, что Q_R и Q_E независимы, $Q_E \sim \chi^{\mathbf{2}}(n-2)$,

$$Q_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2 = \hat{\beta}^2[x,x].$$

Если $\beta = 0$, тогда, согласно (8), $\hat{\beta} \sim \mathcal{N}(0; \sigma/\sqrt{[x,x]})$, поэтому

$$\xi = \hat{\beta}\sqrt{[x,x]}/\sigma \sim \mathcal{N}(0;1), \quad Q_R/\sigma^2 = \xi^2 \sim \chi^{\mathbf{2}}(1).$$

Статистика $T = \sqrt{F}$ имеет распределение Стьюдента с $n-2$ степенями свободы. Она выражается через коэффициент детерминации и, соответ-

ственно, через коэффициент корреляции.

$$T = \sqrt{F} = \sqrt{\frac{Q_R}{Q_E}}\sqrt{n-2} = \frac{\sqrt{Q_R}}{\sqrt{Q-Q_R}}\sqrt{n-2} = \\ = \frac{\sqrt{Q_R/Q}}{\sqrt{1-Q_R/Q}}\sqrt{n-2}.$$

Таким образом получаем статистику для проверки значимости коэффициента корреляции

$$T = \frac{\hat{\varrho}}{\sqrt{1-\hat{\varrho}^2}}\sqrt{n-2} \ \sim \ \mathbf{T}(n-2), \tag{21}$$

которая имеет распределение Стьюдента с $n-2$ степенями свободы в случае справедливости нулевой гипотезы $H_0 : \varrho = 0$.

1.6. Непараметрические коэффициенты корреляции

По выборкам небольшого объема или в случае неоднородных данных при помощи выборочного коэффициента корреляции не всегда удается адекватно представить структуру зависимости между признаками.

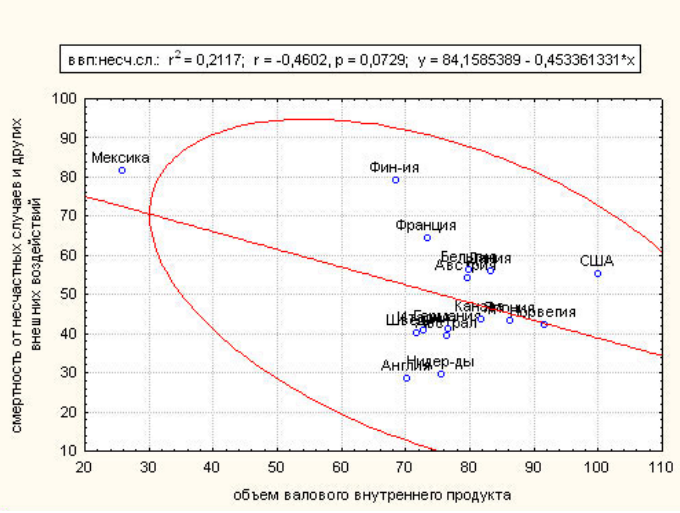

Рис. 1. Наблюдаемая отрицательная зависимость между объемом валового внутреннего продукта и смертностью от несчастных случаев во второй группе стран.

страна	группа	см-ть от несч.сл.	ВВП97	ест.пр.97
Россия	1	187,4	24,2	-5,2
Австрал	2	39,4	76,4	6,7
Австрия	2	54,3	79,5	0,5
Бельгия	2	56,3	79,8	1,2
Болгария	1	66,8	17,9	-7
Англия	2	28,5	70,2	1,6
Венгрия	1	102,4	33,5	-3,8
Германия	2	41,2	76,6	-0,8
Дания	2	55,9	83,2	1,5
Италия	2	40,8	72,8	-0,4
Канада	2	43,6	81,7	4,9
Мексика	2	81,8	25,8	23,4
Нидер-ды	2	29,8	75,6	3,5
Норвегия	2	42,4	91,6	3,5
Польша	1	77	24,7	0,9
Румыния	1	80,8	23,7	-1,9
США	0	55,5	100	5,7
Фин-ия	2	79,1	68,5	1,8
Франция	2	64,5	73,3	3,3
Швеция	2	40,3	71,6	-0,4
Япония	2	43,3	86,2	2,1

Таблица 1. Экономико-демографические данные 1997г.

В этом случае используют непараметрические методы исследования корреляции: коэффициент Спирмена опирается на коррелированность рангов, а коэффициент Кендалла на синхронизированность в отношениях порядков.

Ранговая корреляция Спирмена

Пусть $x'_1, \ldots, x'_n$ и $y'_1, \ldots, y'_n$ – ранги индивидов по признакам X и Y,

- $k_i = x'_i - 1$ — количество индивидов, которых превосходит x_i,
- $l_i = n - x'_i$ — количество индивидов, которые превосходят x_i,

$x_{i*} = k_i - l_i = (x'_i - 1) - (n - x'_i) = 2x'_i - n - 1$. Аналогично построим y_{i*}. Формально эти характеристики выразим как

$$x_{i*} = 2\left(x'_i - \frac{n+1}{2}\right), \qquad y_{i*} = 2\left(y'_i - \frac{n+1}{2}\right).$$

При взаимосвязанности признаков большему значению X соответствует большее значение Y, меньшему меньшее и так далее, в идеальном виде x_{i*} и y_{i*} совпадают, и характеристика

$$Q_{xy} = \sum_{i=1}^{n} x_{i*} y_{i*}$$

принимает наибольшее значение, равное

$$Q = Q_{xx} = Q_{yy} = \sum_{i=1}^{n} x_{i*}^2 = \sum_{i=1}^{n} y_{i*}^2.$$

Если взаимосвязь признаков обратная, то есть меньшему значению X соответствует большее значение признака Y и так далее, то $x_{i*} = -y_{i*}$, и $Q_{xy} = -Q$. Таким образом можно ввести характеристику

$$R_S = \frac{Q_{xy}}{Q}, \tag{22}$$

которая называется *коэффициентом корреляции Спирмена*. Если между признаками существует прямая зависимость вида

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ 1 & 2 & \dots & n-1 & n \end{pmatrix},$$

то коэффиент Спирмена равен единице $R_S = -1$, если имеет место обратная зависимость

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ n & n-1 & \dots & 2 & 1 \end{pmatrix}.$$

Поскольку $\frac{n+1}{2}$ представляет собой средний ранг, коэффициент корреляции Спирмена есть не что иное, как коэффицент корреляции между рангами. Значимость коэффициента Спирмена, как и обычного коэффициента корреляции, проверяется при помощи статистики

$$T = \frac{R_S\sqrt{n-2}}{\sqrt{1-R_S^2}}, \tag{23}$$

которая при некоррелированности рангов X' и Y' имеет распределение Стьюдента с $n-2$ степенями свободы.

ПРИМЕР (продолжение). Коэффициент корреляции Спирмена между объемом ВВП в 1997 году и смертностью от несчастных случаев во второй группе стран равен $R_S = 0.03$, $P = 0.91$. Поскольку корреляция не значима, то наблюдаемое уменьшение смертности при увеличении объема ВВП является не систематическим, а случайным за счет одной страны (Мексики).

Коэффициент корреляция Кендалла

Рассмотрим две выборки $x_1, \dots, x_n$ и $y_1, \dots, y_n$ наблюдений двух признаков, зависимость между которыми нужно как-то измерить. Представим

структуру отношений порядков внутри этих выборок в виде элементов матриц порядка n

$$x_{ik} = \begin{cases} 1, & x_i < x_k, \\ 0, & x_i = x_k, \\ -1, & x_i > x_k \end{cases} \quad \text{и} \quad y_{ik} = \begin{cases} 1, & y_i < y_k, \\ 0, & y_i = y_k, \\ -1, & y_i > y_k \end{cases}$$

и введем показатель однонаправленности неравенств

$$S = \sum_{i=1}^{n} \sum_{k=i+1}^{n} x_{ik} y_{ik} \,.$$

Поскольку число всевозможных пар равно $N = C_n^2 = \frac{n(n-1)}{2}$, то доля однонаправленных пар, называемая *коэффициентом корреляции Кендалла*, равна $T = \frac{S}{N}$.

Например, имеются пять индивидов, которые упорядочены по признаку X (рост), а по признаку Y (вес) выставлены ранги,

$$\begin{matrix} X \\ Y \end{matrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 1 & 3 & 2 \end{pmatrix} .$$

Составим таблицу для вычисления S.

ik	12	13	14	15	23	24	25	34	35	45
x_{ik}	1	1	1	1	1	1	1	1	1	1
y_{ik}	-1	-1	-1	-1	-1	-1	-1	1	1	-1
$x_{ik}y_{ik}$	-1	-1	-1	-1	-1	-1	-1	1	1	-1

Если обозначить через $S_+ = 2$ и $S_- = 8$ количество 1 и -1 соответственно, тогда однонаправленность можно выразить через их разность, $S = 1 \cdot S_+ + (-1) \cdot S_- = S_+ - S_-$. Можно заметить, что S_+ равен количеству элементов выборки Y, больших y_m по всем m при упорядочивании их по признаку X. Учитывая то, что $S_+ + S_- = N$, получаем

$$S = S_+ - S_- = S_+ - (N - S_+) = 2S_+ - N,$$
$$T = \frac{S}{N} = \frac{2S_+ - N}{N} = \frac{2S_+}{N} - 1 \,.$$

Для проверки значимости отклонения от нуля коэффициента корреляции Кендалла используется то, что при независимости признаков $T \sim \mathcal{N}(0, \sigma)$, где $\sigma^2 = \frac{2(2n+5)}{9n(n-1)}$.