1. Основные понятия

После проведения корреляционного анализа, когда выявлено наличие статистически значимых связей между переменными и оценена степень их тесноты, обычно переходят к математическому описанию вида зависимостей с использованием методов регрессионного анализа. С этой целью подбирают класс функций, связывающий результативный показатель y и аргументы $x_1,x_2,...,x_k$, вычисляют оценки параметров уравнения связи и анализируют точность полученного уравнения.

Функция $f(x_1,x_2,...,x_k)$, описывающая зависимость условного среднего значения результативного признака y от заданных значений аргументов, называется *уравнением регрессии.*

Термин “регрессия” (лат. - “regression” - отступление, возврат к чему-либо) введен английским психологом и антропологом Ф. Гальтоном и связан с одним из его первых примеров, в котором Ф. Гальтон, обрабатывая статистические данные, связанные с вопросом о наследственности роста, нашел, что если отцы отклоняются от среднего роста всех отцов на x дюймов, то их сыновья отклоняются от среднего роста всех сыновей меньше, чем на x дюймов. Выявленная тенденция была названа **“регрессией к среднему состоянию”**.

Термин “регрессия” широко используется в статистической литературе, хотя во многих случаях, он недостаточно точно характеризует статистическую зависимость.

Для точного описания уравнения регрессии необходимо знать условный закон распределения результативного показателя y. В статистической практике такую информацию получить обычно не удается, поэтому ограничиваются поиском подходящих аппроксимаций для функции $f(x_1,x_2,...,x_k)$, основанных на предварительном содержательном анализе явления или на исходных статистических данных.

В рамках отдельных модельных допущений о типе распределения вектора показателей $(y, x_1,x_2,...,x_k)$ может быть получен общий вид *уравнения регрессии* $f(X)=M(y/\ X)$, где $X=(x_1,x_2,...,x_k)^T$. Например, в предложении, что исследуемая совокупность показателей подчиняется $(k+1)$ - мерному нормальному закону распределения с вектором математических ожиданий

$$M = \begin{pmatrix} My \\ MX \end{pmatrix},$$

где $MX = \begin{pmatrix} Mx_1 \\ . \\ . \\ . \\ Mx_k \end{pmatrix}$, $\mu_y = My$

и ковариационной матрицей $\sum = \begin{pmatrix} \sigma_{yy} & \sum_{yx}^T \\ \sum_{yx} & \sum_{xx} \end{pmatrix}$,

где $\sigma_{yy} = \sigma_y^2 = M(y - My)^2$ - дисперсия y;

$$\sum_{yx} = \begin{pmatrix} \sigma_{y1} \\ \sigma_{y2} \\ . \\ . \\ . \\ \sigma_{yk} \end{pmatrix}; \quad \sum_{xx} = \begin{pmatrix} \sigma_{11} & \sigma_2 & . & . & . & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & . & . & . & \sigma_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{k1} & \sigma_{k2} & . & . & . & \sigma_{kk} \end{pmatrix};$$

$$\sigma_{ij} = M(x_i - Mx_i)(x_j - Mx_j); \quad \sigma_{jj} = \sigma_j^2 = M(x_j - Mx_j)^2$$

σ_{ij} - ковариация между величинами x_i и x_j, а $\sigma_{jj} = \sigma_j^2$ - дисперсия x_j.

Уравнение регрессии (условное математическое ожидание) имеет вид:

$$M(y / x) = \mu_y + \sum_{yx}^T \cdot \sum_{xx}^{-1} (x - Mx).$$

Таким образом, если многомерная случайная величина (y, $x_1, x_2, ..., x_k$) подчиняется (κ+1)-мерному нормальному закону распределения, то уравнение регрессии результативного показателя y по объясняющим переменным $x_1, x_2, ..., x_k$ имеет линейный по x вид.

Однако в статистической практике обычно приходится ограничиваться поиском подходящих аппроксимаций для неизвестной истинной функции регрессии $f(x)$, так как исследователь не располагает точным знанием условного закона распределения вероятностей анализируемого результативного показателя y при заданных значениях аргументов x.

Рассмотрим взаимоотношение между истиной $f(x)=M(y/x)$, модельной $\tilde{y}$ и оценкой $\hat{y}$ регрессии [1, 26]. Пусть результативный показатель y связан с аргументом x соотношением:

$$y = 2x^{1,5} + \varepsilon ,$$

где ε - случайная величина, имеющая нормальный закон распределения,

причем $M\varepsilon=0$ и $D\varepsilon = \sigma^2$. Истинная функция регрессии в этом случае имеет вид:

$$f(x)=M(y/x)= 2x^{1,5} .$$

Предположим, что точный вид истинного уравнения регрессии нам не известен, но мы располагаем девятью наблюдениями над двумерной случайной величиной, связанной соотношением $y_i = 2x_i^{1,5} + \varepsilon_i$ и представленной на рис. 1.

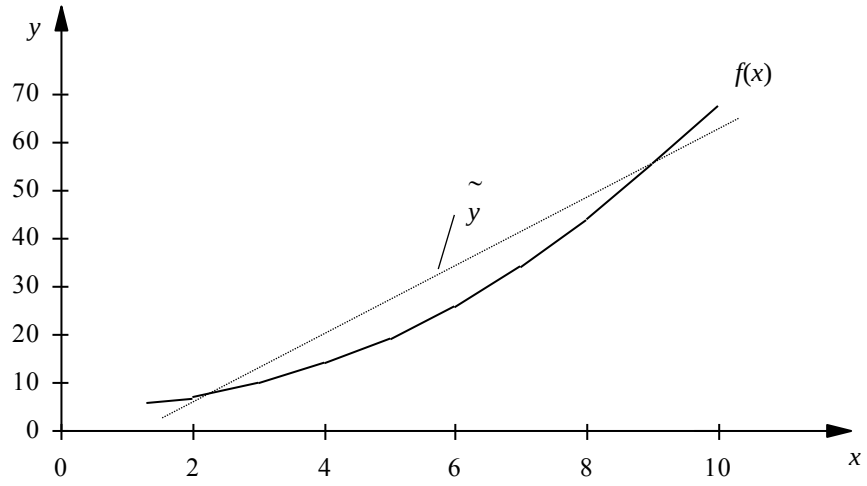

Рис. 1. **Взаимное расположение истинной *f*(*x*) и теоретической $\tilde{y}$ модели регрессии.**

Расположение точек на рис. 1 позволяет ограничиться классом линейных зависимостей вида: $\tilde{y} = \beta_0 + \beta_1 x$.

С помощью метода наименьших квадратов найдем оценку уравнения регрессии.

Для сравнения на рис.1 приводятся графики истинной функции регрессии $f(x) = 2x^{1,5}$ и теоретической аппроксимирующей функции регрессии $\tilde{y} = \beta_0 + \beta_1 x$. К последней сходится по вероятности оценка уравнения регрессии $\hat{y}$ при неограниченном увеличении объема выборки ($n \to \infty$).

Поскольку мы ошиблись в выборе класса функции регрессии, что, к сожалению, достаточно часто встречается в практике статистических исследований, то наши статистические выводы и оценки не будут обладать свойством состоятельности, т.е., как бы мы ни увеличивали объем наблюдений, наша выборочная оценка $\hat{y}$ не будет сходиться к истинной функции регрессии $f(x)$.

Если бы мы правильно выбрали класс функций регрессии, то неточность в описании $f(x)$ с помощью $\hat{y}$ объяснялась бы только ограниченностью выборки и, следовательно, она могла бы быть сделана сколько угодно малой при $n \to \infty$.

С целью наилучшего восстановления по исходным статистическим данным условного значения результативного показателя $y(x)$ и неизвестной функции регрессии $f(x)$=M(y/x) наиболее часто используют следующие *критерии адекватности,* функции потерь [26].

1. *Метод наименьших квадратов*, согласно которому минимизируется квадрат отклонения наблюдаемых значений результативного показателя $y_i (i = 1,2,...,n)$ от модельных значений $\tilde{y}_i = f(x_i, \beta)$, где $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ - коэффициенты уравнения регрессии, x_i - значения вектора аргументов в i-м наблюдении:

$$\sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \to \min_{\beta}.$$

Решается задача отыскания оценки $\hat{\beta}$ вектора β. Получаемая регрессия называется *среднеквадратической*.

2. *Метод наименьших модулей*, согласно которому минимизируется сумма абсолютных отклонений наблюдаемых значений результативного показателя от модульных значений $\tilde{y}_i = f(x_i, \beta)$, т. е.

$$\sum_{i=1}^{n} \left| y_i - f(x_i, \beta) \right| \to \min_{\beta}.$$

Получаемая регрессия называется *среднеабсолютной (медианной).*

3. *Метод минимакса* сводится к минимизации максимума модуля отклонения наблюдаемого значения результативного показателя y_i от модельного значения $f(x_i, \beta)$, т. е.

$$\max_{1 \le i \le n} \left| y_i - f(x_i, \beta) \right| \to \min_{\beta}.$$

Получаемая при этом регрессия называется *минимаксной*.

В практических приложениях часто встречаются задачи, в которых изучается случайная величина *y*, зависящая от некоторого множества переменных $x_1, x_2, \ldots, x_k$ и неизвестных параметров β_j (j=0,1,2,...,$к$). Будем рассматривать (y, $x_1, x_2, \ldots, x_k$) как (k+1)-мерную генеральную совокупность, из которой взята случайная выборка объемов n, где ($y_i, x_{i1}, x_{i2}, \ldots, x_{ik}$) результат i-го наблюдения i=1,2,...,n. Требуется по результатам наблюдений оценить неизвестные параметры β_j (j=0,1,2,...,$к$). Описанная выше задача относится к задачам регрессионного анализа.

Регрессионным анализом называют метод статистического анализа зависимости случайной величины y от переменных x_j *(j= 1,2,...,к), рассматриваемых в регрессионном анализе как неслучайные величины, независимо от истинного закона распределения* x_j.

Часто предполагается, что случайная величина *y* имеет нормальный закон распределения с условным математическим ожиданием $\tilde{y}$, являющимся функцией от аргументов x_j (j= 1,2,...,$к$) и постоянной, не зависящей от аргументов дисперсий σ^2.

Следует помнить, что требование нормальности закона распределения *y* необходимо лишь для проверки значимости уравнения регрессии и его параметров β_j, а также для интервального оценивания β_j. Для получения точечных оценок β_j (j=0,1,2,...,$к$) этого условия не требуется.

В общем виде *линейная модель регрессии* имеет вид:

$$y = \sum_{j=0}^{k} \beta_j \varphi_j(x_1, x_2, ..., x_k) + \varepsilon,$$

где φ_j - некоторая функция его переменных $x_1, x_2, \ldots, x_k$, ε - случайная величина с нулевым математическим ожиданием и дисперсией σ^2.

В регрессионном анализе под *линейной моделью* подразумевают модель, линейно зависящую от неизвестных параметров β_j.

Простейшей линейной будем называть модель, линейно зависящую как от параметров β_j, так и от переменных x_j.

2. Двухмерная линейная модель регрессии

Пусть на основании анализа исследуемого явления предполагается, что в “среднем” *y* линейно зависит от *x*, т.е. имеет место уравнение регрессии.

$$\tilde{y} = M(y/x) = \beta_0 + \beta_1 x\,, \qquad (1)$$

где *M*(*y*/*x*) - условное математическое ожидание случайной величины у при заданном *x*. Объясняющая переменная *x* рассматривается как не случайная величина;

β_0 и β_1 - неизвестные параметры генеральной совокупности, которые подлежат оценке по результатам выборочных наблюдений. С этой целью из двухмерной генеральной совокупности (*x*, *y*) взята выборка объемом n, где (x_i, y_i) результат *i*-го наблюдения (*i*=1,2,...,*n*).

В этом случае линейная регрессионная модель имеет вид:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_1 \quad , \qquad (2)$$

где ε_i - взаимно независимые случайные величины с нулевым математическим ожиданием и дисперсией σ^2, т.е. $M\varepsilon_i = 0$; $D\varepsilon_i = M\varepsilon_i^2 = \sigma^2$ для всех *i*=1,2,...,*n* и

$$M(\varepsilon_i\,\varepsilon_j) = \begin{cases} \sigma^2 \text{ при } i = j \text{ – условие гомоскедастичности, постоянства остаточной дисперсии.} \\ 0 \text{ при } i \neq j \text{ - условие взаимной некоррелировансти регрессионных остатков.} \end{cases}$$

Наша задача – по выборке найти оценки параметров *регрессионной модели.*

2.1. Оценивание параметров регрессии

Согласно методу наименьших квадратов в качестве оценок неизвестных параметров β_0 и β_1 следует брать такие значения выборочных характеристик b_0 и b_1, которые минимизируют сумму квадратов отклонений значений результативного признака y_i от условного математического ожидания $\tilde{y}_i$, т.е.

$$Q = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i x_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \,. \qquad (3)$$

Так как Q дифференцируема по β_0 и β_1, то для отыскания минимума функции (3) найдем частные производные по β_0 и β_1:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1) \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum\limits_{i=1`}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \end{cases}. \qquad (4)$$

Приравняв производные нулю и подставив в (4) вместо β_0 и β_1 их оценки b_0 и b_1, получим:

$$\begin{cases} \sum\limits_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \\ \sum\limits_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i = 0 \end{cases} \text{ или } \begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i \end{cases}. \qquad (5)$$

Данная система уравнений называется *системой нормальных уравнений.*

Решая систему (5) относительно b_0 и b_1, получим

$$b_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - \frac{1}{n}\sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{\sum\limits_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum\limits_{i=1}^{n} x_i)^2}; \; b_0 = \frac{1}{n}\sum\limits_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum\limits_{i=1}^{n} x_i \qquad (6)$$

Перейдя к средним, будем иметь

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}; \;\; b_0 = \bar{y} - b_1 \bar{x} \qquad (6')$$

Таким образом, имеем оценку уравнения регрессии:

$$\hat{y} = b_0 + b_1 x. \qquad (7)$$

Докажем, что в случае нормального закона распределения случайной величины ε_i, а отсюда и y_i, согласно (2), оценки метода наименьших квадратов и максимального правдоподобия совпадают.

Пусть из двумерной генеральной совокупности (x,y) взята независимая выборка (x_i, y_i), где i=1,2,...,n, объемом n.

Будем рассматривать y_i как независимые нормальные случайные величины с математическим ожиданием $\tilde{y}_i = M(y_i / x_i)$, являющимся функцией от x_i согласно (1), и постоянной дисперсией σ^2.

Тогда $f(y_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i-\tilde{y}_i)^2}{2\sigma^2}}$, где $\tilde{y}_i = \beta_0 + \beta_1 x_i$, и функция правдоподобия примет вид с учетом независимости наблюдения:

$$L = P(y_1, x_1; y_2, x_2; \ldots; y_n, x_n; \beta_0, \beta_1; \sigma^2) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(y_i-\tilde{y}_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} e^{-\frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{2\sigma^2}}.$$

Согласно *методу наибольшего правдоподобия* в качестве оценок параметров β_0, β_1 и σ^2 возьмем значения b_0, b_1 и s^2, максимизирующие функцию правдоподобия L. При заданных $x_1, x_2, \ldots, x_n$ и постоянном σ^2 функция правдоподобия L достигнет максимума, когда показатель степени при e будет минимальным, т.е. при условии минимума функции $Q = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, что совпадает с условием (3) нахождения оценок b_0, b_1 по методу наименьших квадратов. Таким образом, оценки b_0, b_1 обладают свойствами оценок наибольшего правдоподобия.

Однако функция правдоподобия L зависит также и от параметра σ. Из условия $\frac{\partial L}{\partial \sigma} = 0$ найдем оценку s^2 наибольшего правдоподобия параметра σ^2:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2. \qquad (8)$$

Несмещенная оценка параметра σ^2 равна:

$$\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2. \qquad (9)$$

Исследуем свойства оценок b_0 и b_1.

2.2. Определение интервальной оценки для β_0

Будем рассматривать модель регрессионного анализа вида:

$$y_i = \beta_0 + \beta_1 x_i' + \varepsilon_i \quad \text{или} \quad y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i \tag{10}$$

где $x_i' = (x_i - \bar{x})$ - центрированные величины, удовлетворяющие условию

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 .$$

Тогда оценки b_0 и b_1 метода наименьших квадратов согласно (6) равны:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad b_0 = \frac{1}{n} \sum_{i=1}^{n} y_i . \tag{11}$$

Учитывая (10), получим $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} [\beta_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i] = n\beta_0 + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) + \sum_{i=1}^{n} \varepsilon_i$.

Тогда, с учетом (11), будем иметь $$b_0 = \beta_0 + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i . \tag{12}$$

Величина b_0 есть линейная функция нормальных случайных величин ε_i. Следовательно, она также имеет нормальный закон распределения с математическим ожиданием:

$$Mb_0 = M(\beta_0 + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i) = \beta_0 + \frac{1}{n} \sum_{i=1}^{n} M\varepsilon_i = \beta_0 , \tag{13}$$

так как по условию $M\varepsilon_i = 0$. Дисперсия оценки b_0 равна :

$$Db_0 = D(\beta_0 + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i) = D(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i) = \frac{1}{n^2} \sum_{i=1}^{n} D\varepsilon_i = \frac{\sigma^2}{n} . \tag{14}$$

Здесь учитывалось, что ε_i взаимно независимые случайные величины с дисперсией $D\varepsilon_i = \sigma^2$ для всех i=1,2,...,n. Подставляя вместо σ^2 несмещенную оценку $\hat{s}^2$, получим оценку дисперсии $\hat{s}_{b_0}^2$, для b_0, $\hat{s}_{b0}^2 = \frac{\hat{S}^2}{n}$.

Таким образом, b_0 есть случайная величина, имеющая нормальный закон распределения $b_0 \in N(\beta_0; \frac{\sigma}{\sqrt{n}})$.

Отсюда следует, что величина $$z = \frac{b_0 - \beta_0}{\sigma} \sqrt{n} \in N(0,1) \tag{15}$$

имеет нормированный нормальный закон распределения.

С другой стороны, статистика

$$u^2 = \frac{ns^2}{\sigma^2} \in \chi^2(\nu = n-2) \tag{16}$$

имеет χ^2-распределение с $\nu = n-2$ степенями свободы, так как уравнение регрессии определяется двумя параметрами b_0 и b_1, которые подлежат оцениванию.

Отсюда следует, что статистика

$$t = \frac{z}{u}\sqrt{\nu} = \frac{b_0 - \beta_0}{\sigma}\sqrt{n} \cdot \frac{\sigma}{\sqrt{n}\cdot s}\sqrt{n-2} = \frac{b_0 - \beta_0}{s}\sqrt{n-2} \in St(\nu = n-2)$$

имеет *t*-распределение Стьюдента с $\nu = n-2$ степенями свободы.

С помощью статистики *t* построим с доверительной вероятностью γ интервальную оценку для β_0 из условия:

$$P\left\{-t_\gamma \le \frac{b_0 - \beta_0}{s}\sqrt{n-2} \le t_\gamma\right\} = \gamma\,.$$

Откуда, решая неравенства относительно β_0 получим

$$\beta_0 \in \left[b_0 \pm t_\gamma \frac{s}{\sqrt{n-2}}\right] \tag{17}$$

или, учитывая, что $\hat{s}_{b_0} = \frac{\hat{s}}{\sqrt{n}} = \frac{s}{\sqrt{n-2}}$, будем иметь: $\beta_0 \in \left[b_0 \pm t_\gamma \hat{s}_{b_0}\right]$, где t_γ определяется по таблице распределения Стьюдента (*t*-распределение) для уровней значимости $\alpha = 1-\gamma$ и числа степеней свободы $\nu = n-2$. Выражение (17) показывает, что β_0 принадлежит интервалу, границы которого заданы в квадратных скобках.

2.3. Определение интервальной оценки и проверка значимости β_1

С учетом (10) рассмотрим выражение

$$\sum_{i=1}^{n}(x_i-\bar{x})y_i=\sum_{i=1}^{n}(x_i-\bar{x})[\beta_0+\beta_1(x_0-\bar{x})+\varepsilon_i]=$$

$$=\beta_0\sum_{i=1}^{n}(x_i-\bar{x})+\beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i=$$

$$=\beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i.$$

При этом учитывалось, что $\sum_{i=1}^{n}(x_i-\bar{x})=0$. Решив уравнение относительно β_1, получим:

$$\beta_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\frac{\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

а с учетом (11) будем иметь:

$$b_1=\beta_1+\frac{\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \qquad (18)$$

Из (18) следует, что b_1 есть линейная функция независимых нормально распределенных случайных величин $\varepsilon_i\in N(0,\sigma)$, где i=1,2,...,n. Следовательно, она также имеет нормальный закон распределения. Определим математическое ожидание и дисперсию b_1.

Учитывая, что математическое ожидание суммы равно сумме математических ожиданий, что неслучайный множитель $(x_i-\bar{x})$ можно вынести за знак математического ожидания и $M\varepsilon_i=0$, получим:

$$Mb_1=M\beta_1+\frac{\sum_{i=1}^{n}(x_i-\bar{x})M\varepsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\beta_1. \qquad (19)$$

Так как ε_i есть независимые между собой случайные величины с дисперсией $D\varepsilon_i=\sigma^2$, а дисперсия постоянной величины равна нулю, т.е. $D\beta_1=0$, то

$$Db_1=D\beta_1+\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2D\varepsilon_i}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2},$$

откуда получим

$$Db_1 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{20}$$

Мы доказали, что в b_1 есть случайная величина, имеющая нормальный закон распределения:

$$b_1 \in N\left(\beta_1; \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right).$$

Отсюда следует, что

$$z = \frac{b_1 - \beta_1}{\sigma}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \in N(0,1). \tag{21}$$

Учитывая независимость случайных величин (16) и (21), получим статистику, имеющую *t*-распределение с $\nu = n - 2$ степенями свободы:

$$\boldsymbol{t = \frac{z}{u}\sqrt{n-2} = \frac{(b_1 - \beta_1)\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{s}\sqrt{\frac{n-2}{n}};}$$
$$\boldsymbol{t = \frac{(b_1 - \beta_1)}{\hat{s}}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \in St(\nu = n-2),} \tag{22}$$

где $\hat{s} = s\sqrt{\frac{n}{n-2}}$.

Интервальную оценку для β_1 с надежностью γ найдем из условия: $P(-t_\gamma \le t \le t_\gamma) = \gamma$.

После преобразования с учетом (22) получим

$$\beta_1 \in \left[b_1 \pm t_\gamma \frac{\hat{s}}{\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2}}\right]$$

или

$$\beta_1 \in \left[b_1 \pm t_\gamma \hat{s}_{b1}\right], \tag{23}$$

где t_γ - находят по таблице *t*-распределения при $\alpha = 1 - \gamma$ и $\nu = n - 2$;

$\hat{s}^2_{b_1} = \frac{\hat{s}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ - несмещенная оценка дисперсии D_{b_1};

$\hat{s}_{b_1}$ - оценка среднего квадратического отклонения величины b_1.

С надежностью γ найдем интервальную оценку для σ^2 с помощью статистики (16):

$$\sigma^2 \in [\frac{ns^2}{\chi_1^2}; \frac{ns^2}{\chi_2^2}], \tag{24}$$

где χ_1^2, χ_2^2 находят по таблице χ^2 -распределения для числа степеней свободы $\nu = n-2$ и вероятностей соответственно $\frac{1+\gamma}{2}$ и $\frac{1-\gamma}{2}$.

Установление значимости простейшего линейного уравнения регрессии $\tilde{y} = \beta_0 + \beta_1 x$ сводится к проверке при заданном α нулевой гипотезы о значимости коэффициента регрессии β_1, т.е. гипотезы $H_0: \beta_1 = 0$ при альтернативной гипотезе $H_1: \beta_1 \neq 0$.

С этой целью используется *t*-критерий и значение статистики критерия

$$t_1 = \frac{b_1}{\hat{s}_{b_1}} \tag{25}$$

сравнивают с критическим значением $t_{кр}(\alpha; \nu = n-2)$, найденным при заданном α и $\nu = n-2$ по таблице *t*-распределения.

Гипотеза $H_0: \beta_1 = 0$ отвергается с вероятностью ошибки α при выполнении неравенства $|t_1| > t_{кр}\ (\alpha, \nu = n-2)$ и уравнение регрессии считается значимым. В противном случае, т.е. если $|t_1| \leq t_{кр}$, гипотеза $H_0: \beta_1 = 0$ не отвергается и уравнение регрессии считают незначимым и на этом регрессионный анализ заканчивается.

Для значимого уравнения регрессии представляет интерес построение интервальных оценок для коэффициента регрессии β_1, свободного члена β_0 и самого уравнения $\tilde{y}$.

2.4. Определение интервальной оценки для условного математического ожидания

Пусть имеем уравнение регрессии:

$$\tilde{y} = \beta_0 + \beta_1 (x - \bar{x}) \tag{26}$$

и его оценку: $\hat{y} = b_0 + b_1 (x - \bar{x})$, где b_0, b_1 - оценки метода наименьших квадратов параметров уравнения β_0, β_1 .

Величина $\hat{y}$ есть линейная функция двух случайных величин b_0 и b_1, имеющих нормальный закон распределения. Следовательно, $\hat{y}$ также имеет нормальный закон распределения. Параметры этого закона получим, учитывая выражения (12) и (19):

$$M\hat{y} = M\left[b_0 + b_1(x - \bar{x})\right] = Mb_0 + (x - \bar{x})Mb_1 .$$

Откуда $M\hat{y} = \beta_0 + \beta_1(x - \bar{x}) = \tilde{y}$. Для определения дисперсии $D\hat{y}$ предварительно докажем независимость величин b_0 и b_1.

Так как величины b_0 и b_1 имеют нормальный закон распределения, то независимость этих величин следует из их некоррелированности. Следовательно, нам достаточно доказать, что $M(b_0 - \beta_0)(b_1 - \beta_1) = 0$.

Учитывая выражения (12) и (18) и, что x_i есть неслучайная величина, получим:

$$M(b_0 - \beta_0)(b_1 - \beta_1) = M\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)\left(\frac{\sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) =$$

$$= \frac{1}{n\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]}\sum_{i=1}^{n} M\varepsilon_i \sum_{j=1}^{n}(x_j - \bar{x})\varepsilon_j .$$

Так как ε_i (i= 1, 2, ...,n) по условию есть независимые случайные величины с $M\varepsilon_i$=0, то $M\varepsilon_i\varepsilon_j = 0$ при $i \neq j$, где i, j=1,2,...,n. Следовательно,

$$M\varepsilon_i \sum_{j=1}^{n}(x_i - \bar{x})\varepsilon_j = (x_i - \bar{x})M\varepsilon_i^2 = (x_i - \bar{x})\sigma^2 ,$$

где $M\varepsilon_i^2 = \sigma^2$. Учитывая, что $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, после подстановки окончательно получим:

$$M(b_0 - \beta_0)(b_1 - \beta_1) = \frac{1}{n\sum_{i=1}^{n}(x_1 - \bar{x})^2}\sigma^2\sum_{i=1}^{n}(x_i - \bar{x}) = 0 .$$

Этот результат получен для центрированных величин $(x_i - \bar{x})$, для которых выполняется условие $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. В этом случае b_0 и b_1 независимые случайные величины. Тогда

согласно выражению (26) дисперсия величины $\hat{y}$ равна сумме дисперсий слагаемых, т.е.: $D\hat{y} = Db_0 + (x - \bar{x})^2 Db_1$.

получим:

$$D\hat{y} = \frac{\sigma^2}{n}\left(1 + n\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Таким образом

$$\hat{y} \in N\left(\tilde{y}; \frac{\sigma}{\sqrt{n}}\sqrt{1 + n\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right). \quad (27)$$

Тогда нормированный нормальный закон распределения имеет величина:

$$z = \frac{\hat{y} - \tilde{y}}{\frac{\sigma}{\sqrt{n}}\sqrt{1 + n\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \in N(0,1). \quad (28)$$

получим выборочную характеристику:

$$t = \frac{\hat{y} - \tilde{y}}{\hat{s}\sqrt{1 + n\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}\sqrt{n-2} \in St(\nu = n-2), \quad (29)$$

которая имеет распределение Стьюдента (*t*-распределение) с $\nu = n - 2$ степенями свободы.

Тогда с надежностью γ доверительный интервал для $\tilde{y}$ при заданном $x = x_0$ равен:

$$\tilde{y}_{x_0} \in \left[(b_0 + b_1 x_0) \pm t_\gamma \hat{s}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right], \quad (30)$$

где t_γ определяется по таблице распределения Стьюдента для уровня значимости $\alpha = 1 - \gamma$ и числа степеней свободы $\nu = n - 2$.

Интервальная оценка для прогнозного значения y в точке x_{n+1} определяется как:

$$y_{n+1} \in \left[(b_0 + b_1 x_{n+1}) \pm t_\gamma \hat{s} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + 1} \right] \qquad (31)$$

Из (31) следует, что y прогнозного значения y_{n+1} дисперсия на $\hat{s}^2$ больше, чем у величины $\hat{y}$ на величину дисперсии. Согласно (30) по мере удаления x_0 от среднего значения ($\bar{x}$) ширина доверительного интервала увеличивается, а точность оценки $\tilde{y}$ снижается. Доверительный интервал имеет наименьшую величину, когда $x_0 = \bar{x}$. Расположение доверительного интервала для $\tilde{y}$, при заданной γ, иллюстрирует рис. 2.

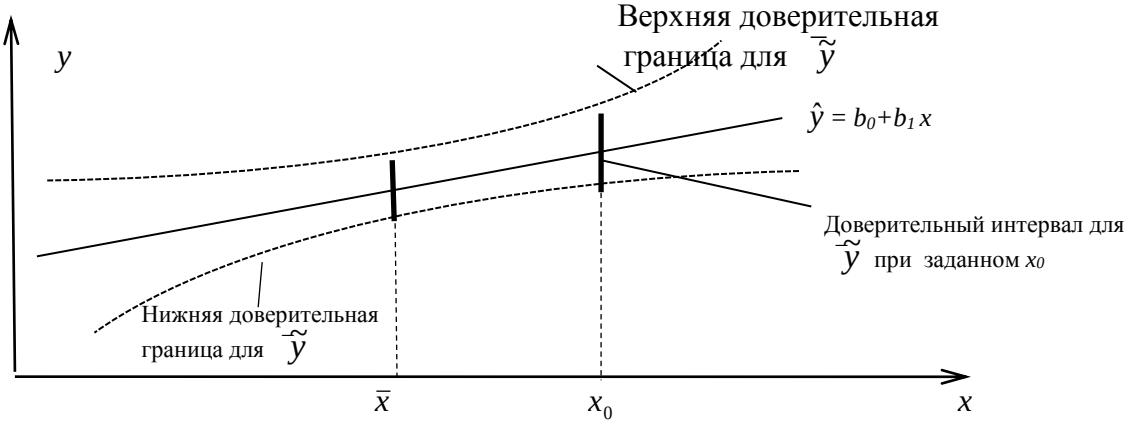

Рис.2. **Расположение доверительных границ в случае линейной регрессии**

2.5. Модель регрессии в случае двумерной нормальной генеральной совокупности

Рассмотрим генеральную совокупность с двумя признаками x и y, совместное распределение которых задано плотностью двумерного нормального закона

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\{Q_2(x,y)\}, \qquad (32)$$

где $Q_2(x,y) = \frac{1}{2\sqrt{1-\rho^2}}[(\frac{x-\mu_x}{\sigma_x})^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y} + (\frac{y-\mu_y}{\sigma_y})^2]$, определяемого пятью параметрами: двумя математическими ожиданиями $M_x = \mu_x$ и $M_y = \mu_y$;

двумя дисперсиями $Dx = \sigma_x^2$ и $Dy = \sigma_y^2$; коэффициентом корреляции

$$M[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}] = \rho, \text{ где } \rho^2 \neq 1.$$

Имея эти параметры, можно получить линейные уравнения регрессии, показывающие изменение условных математических ожиданий одной величины в зависимости от изменения значений соответствующих случайных аргументов:

$My/x - My = \beta_{yx}(x - Mx)$ - линейная регрессия *y* по *x*;

$Mx/y - Mx = \beta_{xy}(y - My)$ - линейная регрессия *x* по *y*;

$\beta_{yx} = \rho\dfrac{\sigma_y}{\sigma_x}$ - коэффициент регрессии *y* на *x*;

$\beta_{xy} = \rho\dfrac{\sigma_x}{\sigma_y}$ - коэффициент регрессии *x* на *y*.

Из этих выражений следует, что *знаки при коэффициентах регрессии и корреляции всегда совпадают* и $\beta_{xy} \cdot \beta_{yx} = \rho^2$.

Квадрат коэффициента корреляции ρ^2 называют коэффициентом детерминации. В рассматриваемой модели он показывает долю дисперсии одной случайной величины, обусловленную вариацией другой.

Коэффициент регрессии β_{yx} показывает, на сколько единиц своего измерения в среднем увеличится (при $\beta>0$) или уменьшится (при $\beta<0$) величина y, т.е. (My/x), если x увеличить на единицу своего измерения.

Задача двумерного регрессионного анализа состоит, прежде всего, в оценке пяти параметров, определяющих генеральную совокупность.

В качестве точечных оценок неизвестных начальных моментов первого и второго порядка генеральной совокупности берутся соответствующие выборочные моменты.

Точечные же оценки других параметров получают как функции от начальных моментов. Таким образом, будем иметь: $\bar{x}$ - оценка для μ_x, $\bar{y}$ - оценка для μ_y, $\bar{x}^2$ - оценка для $M(x^2)$, $\bar{y}^2$ - оценка для $M(y^2)$, $\overline{xy}$ - оценка для $M(xy)$. Откуда:

$s_x^2 = \bar{x}^2 - (\bar{x})^2$ - оценка для σ_x^2, $\qquad s_y^2 = \bar{y}^2 - (\bar{y})^2$ - оценка для σ_y^2,

$r = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x s_y}$ - оценка для ρ.

Оценки генеральных коэффициентов регрессии β_{yx} и β_{xy} получаются соответственно по

формулам:

$$b_{yx} = r\frac{s_y}{s_x}, \qquad b_{xy} = r\frac{s_x}{s_y}, \tag{33}$$

откуда оценки уравнений регрессии имеют вид:

$$\overline{y/x} - \bar{y} = b_{yx}(x - \bar{x}), \qquad \overline{x/y} - \bar{x} = b_{xy}(y - \bar{y}). \tag{34}$$

При этом $\overline{y/x}$ и $\overline{x/y}$ - являются оценками условных математических ожиданий *My*/*x* и *Mx*/*y* генеральной совокупности.

Следует отметить, что вышеприведенные точечные оценки являются состоятельными, а $\bar{x}$ и $\bar{y}$ несмещенными и эффективными. Кроме того, распределение выборочных средних $(\bar{x}, \bar{y})$ не зависит от распределения (s_x^2, s_y^2, r).

На примере двумерного распределения мы показали, что в случае многомерного (*k*-мерного) нормального закона распределения легко прослеживается связь между переменными, характеризирующими тесноту и вид связи в моделях корреляционного и линейного регрессионного анализа.