А.Г. Белов

МНК-ОЦЕНКИ ПАРАМЕТРОВ ЛИНЕЙНОЙ ПАРНОЙ КОНФЛЮЭНТНОЙ МОДЕЛИ

Введение. В различных областях естествознания часто возникает задача исследования априори постулируемых функциональных зависимостей между количественными неслучайными переменными, значения которых определяются из эксперимента. Об экспериментальных ошибках при измерении всех переменных предполагается лишь, что они случайны, некоррелированы и распределены по некоторому вероятностному закону с нулевым математическим ожиданием и конечной постоянной дисперсией. В работе рассматривается часто используемый в такой ситуации метод наименьших квадратов (МНК) для нахождения оценок параметров в случае двухпараметрической линейной функциональной зависимости между изучаемыми переменными с ошибками в обеих переменных (простая линейная конфлюэнтная модель).

Постановка задачи. Предположим, что изучаются n объектов с характерными для каждого из них двумя признаками X,Y, (например, рост и вес для каждого из n человек или доход и расход для каждой из n семей). Обозначим через X_i,Y_i гипотетические неизвестные значения соответствующих признаков X,Y для каждого i-ого объекта, $i=\overline{1,n}$.

Пусть в отношении X и Y сделано предположение, что они связаны линейной зависимостью

$$Y = aX + b, \tag{1}$$

где коэффициенты a,b неизвестны и подлежат определению. Тогда соотношению (1) соответствует связь между значениями признаков для каждого из n объектов, имеющая вид

$$Y_i = aX_i + b, \quad i=\overline{1,n}. \tag{2}$$

Будем считать также, что для каждого из n изучаемых объектов независимо от других объектов одновременно измеряются два интересующих нас признака X,Y. Результаты измерений представляют собой набор $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$, где x_i,y_i – значения признаков X,Y, соответственно, зафиксированных при измерении у i-ого объекта.

Однако при проведении измерений каждого из признаков в общем случае могут допускаться неточности. Тогда справедливы равенства:

$$x_i = X_i + e_{xi}, \quad y_i = Y_i + e_{yi}, \qquad i = \overline{1,n}, \tag{3}$$

где e_{xi}, e_{yi}- ошибки измерений, рассматриваемые как значения некоррелированных случайных величин (с.в.) e_x, e_y с нулевыми математическими ожиданиями и конечными дисперсиями:

$$Ee_x = 0,\ Ee_y = 0,\ De_x = \sigma_x^2,\ De_y = \sigma_y^2,\ \ \sigma_x^2 + \sigma_y^2 \neq 0,\ \operatorname{cov}(e_x, e_y) = 0.$$

Из соотношений (2), (3) получаем следующую «структурную» [1,Гл.29] связь между измерениями

$$y_i = ax_i + b + \left(e_{yi} - ae_{xi}\right), \qquad i = \overline{1,n}. \tag{4}$$

При этом, измерения $\left(x_i, y_i\right), i = \overline{1,n},$ можно рассматривать как результат наблюдений с.в. $\left(x, y\right)$. В силу (3) имеем

$$Ex = X,\ Ey = Y, \tag{5}$$

а в силу (1) справедливо равенство

$$Ey = aEx + b. \tag{6}$$

Возникает следующая задача: в условиях пассивного эксперимента по результатам измерений $\left(x_i, y_i\right), i = \overline{1,n},$ построить статистические точечные оценки неизвестных признаков $\left(X_i, Y_i\right)$ и коэффициентов a, b (2), для которых достигается минимальное значение суммы $\left\{\sum_{i=1}^{n} \frac{e_{xi}^2}{\sigma_x^2} + \sum_{i=1}^{n} \frac{e_{yi}^2}{\sigma_y^2}\right\}$. Выбор последнего функционала связан с тем, что его минимизация равносильна максимизации функции правдоподобия в предположении двумерного нормального закона распределения с.в. $\left(e_x, e_y\right)$ [1, Гл.29].

Ниже будет получен явный вид МНК-оценок для простой линейной конфлюэнтной модели, показана их состоятельность и эквивалентность

оценкам, получаемым по методу максимального правдоподобия при предположении нормальности ошибок [1, Гл.29].

Оценка параметров. В силу условия (2) справедливо равенство

$$\min_{\substack{a,b,X_i,Y_i,i=\overline{1,n}\\ Y_i=aX_i+b}}\left\{\sum_{i=1}^{n}\frac{e_{xi}^2}{\sigma_x^2}+\sum_{i=1}^{n}\frac{e_{yi}^2}{\sigma_y^2}\right\}=\min_{a,b,X_i,i=\overline{1,n}}\left\{\sum_{i=1}^{n}\frac{(x_i-X_i)^2}{\sigma_x^2}+\sum_{i=1}^{n}\frac{(y_i-aX_i-b)^2}{\sigma_y^2}\right\}. \quad (7)$$

Обозначим

$$S=S(a,b,X_1,X_2,...,X_n)=\sum_{i=1}^{n}\frac{(x_i-X_i)^2}{\sigma_x^2}+\sum_{i=1}^{n}\frac{(y_i-aX_i-b)^2}{\sigma_y^2}, \quad (8)$$

тогда задача минимизации (7) сводится к выбору в качестве оценок для a, b и $X_i, i=\overline{1,n}$ таких функций $\hat{a}$, $\hat{b}$ и $\hat{X}_i, i=\overline{1,n}$ от измерений $x_i, y_i, i=\overline{1,n}$, для которых выполнено соотношение

$$S(\hat{a},\hat{b},\hat{X}_1,\hat{X}_2,...,\hat{X}_n)\le S(a,b,X_1,X_2,...,X_n)$$

при любых допустимых значений a, b и $X_i, i=\overline{1,n}$.

Для решения задачи сначала найдем множество всех возможных оценок a, b и $X_i, i=\overline{1,n}$, для которых функция $S(a,b,X_1,X_2,...,X_n)$ может иметь экстремальное значение, а потом из этого множества выберем те оценки, при которых S принимает наименьшее значение.

Утверждение 1. Решения $\hat{a}$, $\hat{b}$ и $\hat{X}_i, i=\overline{1,n}$ задачи (7) удовлетворяют равенствам

$$\begin{cases}\hat{X}_i=\dfrac{\sigma_y^2x_i+\sigma_x^2\hat{a}(y_i-\hat{b})}{\sigma_y^2+\sigma_x^2\hat{a}^2}, i=\overline{1,n.}\\ \hat{a}^2\sigma_x^2s_{xy}-\hat{a}(\sigma_x^2s_y^2-\sigma_y^2s_x^2)-\sigma_y^2s_{xy}=0,\\ \hat{b}=\overline{y}-\hat{a}\overline{x}.\end{cases} \quad (9)$$

где $\overline{x},\overline{y},s_x^2,s_y^2,s_{xy}$ - выборочные средние, дисперсии и ковариация измерений $x_i, y_i, i=\overline{1,n}$, определяемые как

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\ s_x^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2,\ s_y^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2,$$

$$s_{xy}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right).$$

Доказательство. Дифференцируемая функция S (8) может иметь экстремумы только при тех значениях a, b и $X_i, i=\overline{1,n}$, при которых частные производные $S\left(a,b,X_1,X_2,...,X_n\right)$ по всем $n+2$ неизвестным равны нулю. Для их определения имеем следующие системы алгебраических уравнений

$$\begin{cases}\dfrac{\partial S}{\partial X_i}=0,\\ i=\overline{1,n},\\ \dfrac{\partial S}{\partial a}=0,\\ \dfrac{\partial S}{\partial b}=0,\end{cases}\Leftrightarrow\begin{cases}X_i\left(\sigma_y^2+\sigma_x^2a^2\right)=\sigma_y^2x_i+\sigma_x^2\left(ay_i-ab\right),\\ i=\overline{1,n},\\ \sum\limits_{i=1}^{n}\left(y_i-aX_i-b\right)X_i=0,\\ \sum\limits_{i=1}^{n}\left(y_i-aX_i-b\right)=0.\end{cases}$$

Выразим из первого уравнения X_i и подставим в оставшиеся два. После преобразований получим систему уравнений

$$\begin{cases}X_i=\dfrac{\sigma_y^2x_i+\sigma_x^2\left(ay_i-ab\right)}{\sigma_y^2+\sigma_x^2a^2},\\ i=\overline{1,n},\\ a^2\sigma_x^2s_{xy}-a\left(\sigma_x^2s_y^2-\sigma_y^2s_x^2\right)-\sigma_y^2s_{xy}=0,\\ b=\overline{y}-a\overline{x}.\end{cases}$$

Что и требовалось доказать.

Таким образом, задача (7) может быть переформулирована в следующем виде: найти, среди всего множества оценок $\hat{a}$, $\hat{b}$ и $\hat{X}_i, i=\overline{1,n}$, удовлетворяющих (9), те, которые дают наименьшее значение $S\left(a,b,X_1,X_2,...,X_n\right)$. Более того, после подстановки найденных $\hat{X}_i=\dfrac{\sigma_y^2x_i+\sigma_x^2\left(\hat{a}y_i-\hat{a}\hat{b}\right)}{\sigma_y^2+\sigma_x^2\hat{a}^2}$ в $S\left(a,b,X_1,X_2,...,X_n\right)$ и проведения простейших преобразований вместо (7) приходим к следующей задаче:

найти, среди всего множества оценок $\hat{a}$, $\hat{b}$, удовлетворяющих

$$\begin{cases} \hat{a}^2\sigma_x^2 s_{xy} - \hat{a}\left(\sigma_x^2 s_y^2 - \sigma_y^2 s_x^2\right) - \sigma_y^2 s_{xy} = 0, \\ \hat{b} = \overline{y} - \hat{a}\overline{x}, \end{cases} \tag{9'}$$

те, которые дают наименьшее значение функционала

$$\tilde{S}\left(\hat{a},\hat{b}\right) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{a}x_i - \hat{b}\right)^2}{\sigma_y^2 + \hat{a}^2\sigma_x^2}. \tag{8'}$$

Нетрудно показать, что система (9') является системой нормальных уравнений задачи

$$\min_{a,b}\tilde{S}(a,b) = \min_{a,b}\frac{\sum_{i=1}^{n}\left(y_i - ax_i - b\right)^2}{\sigma_y^2 + a^2\sigma_x^2} \tag{7'}$$

минимизации функционала $\tilde{S}(a,b)$ по множеству всех возможных значений параметров a, b.

Действительно, система частных производных задачи (7') имеет вид

$$\begin{cases} \dfrac{\partial\tilde{S}}{\partial a} = \dfrac{-2}{\left(\sigma_y^2 + a^2\sigma_x^2\right)}\left\{\sum_{i=1}^{n}x_i\left(y_i - ax_i - b\right) + \dfrac{a\sigma_x^2\sum_{i=1}^{n}\left(y_i - ax_i - b\right)^2}{\left(\sigma_y^2 + a^2\sigma_x^2\right)}\right\} = 0, \\ \dfrac{\partial\tilde{S}}{\partial b} = \dfrac{-2\sum_{i=1}^{n}\left(y_i - ax_i - b\right)}{\left(\sigma_y^2 + a^2\sigma_x^2\right)} = 0. \end{cases} \tag{9''}$$

После несложных, но трудоемких преобразований получаем (9').

Следствие 1. При $\sigma_x^2 \neq \sigma_y^2$, $\sigma_x^2 \neq 0, \sigma_y^2 \neq 0, s_{xy} \neq 0$ наименьшее значение функции $\tilde{S}$ (8'), а следовательно S (8), задачи (7') ((7)), достигается на оценках, определяемых равенствами

$$\begin{cases} \hat{a} = \dfrac{\left(\sigma_x^2 s_y^2 - \sigma_y^2 s_x^2\right) + \sqrt{\left(\sigma_x^2 s_y^2 - \sigma_y^2 s_x^2\right)^2 + 4\sigma_x^2 \sigma_y^2 s_{xy}^2}}{2\sigma_x^2 s_{xy}} = \\ = \dfrac{\left(\dfrac{s_y}{s_x} - \dfrac{\sigma_y^2 s_x}{\sigma_x^2 s_y}\right) + \sqrt{\left(\dfrac{s_y}{s_x} - \dfrac{\sigma_y^2 s_x}{\sigma_x^2 s_y}\right)^2 + 4\dfrac{\sigma_y^2}{\sigma_x^2} r_{xy}^2}}{2r_{xy}}, \\ \hat{b} = \overline{y} - \hat{a}\overline{x}, \end{cases} \tag{10}$$

где $r_{xy} = \dfrac{s_{xy}}{s_x s_y}$ - выборочный коэффициент корреляции.

Доказательство. Введем обозначения для корней квадратного трехчлена в (9) ((9'))

$$\hat{a}^{(+)} = \frac{\left(\sigma_x^2 s_y^2 - \sigma_y^2 s_x^2\right) + \sqrt{\left(\sigma_x^2 s_y^2 - \sigma_y^2 s_x^2\right)^2 + 4\sigma_x^2 \sigma_y^2 s_{xy}^2}}{2\sigma_x^2 s_{xy}},$$

$$\hat{a}^{(-)} = \frac{\left(\sigma_x^2 s_y^2 - \sigma_y^2 s_x^2\right) - \sqrt{\left(\sigma_x^2 s_y^2 - \sigma_y^2 s_x^2\right)^2 + 4\sigma_x^2 \sigma_y^2 s_{xy}^2}}{2\sigma_x^2 s_{xy}}.$$

Подставив $\hat{b} = \overline{y} - \hat{a}\overline{x}$ в $\tilde{S}\left(\hat{a}, \hat{b}\right)$ получим

$$\hat{S}\left(\hat{a}\right) = \frac{\sum_{i=1}^{n}\left(\left(y_i - \overline{y}\right) - \hat{a}\left(x_i - \overline{x}\right)\right)^2}{\sigma_y^2 + \hat{a}^2\sigma_x^2} = \frac{n s_y^2 + \hat{a}^2 n s_x^2 - 2\hat{a} n s_{xy}}{\sigma_y^2 + \hat{a}^2\sigma_x^2}.$$

Сравним $\hat{S}\left(\hat{a}^{(+)}\right)$ с $\hat{S}\left(\hat{a}^{(-)}\right)$. Рассмотрим 2 случая. Пусть $s_{xy} > 0$. Тогда очевидно $\hat{a}^{(+)} > 0, \hat{a}^{(-)} < 0$ и поэтому $\hat{S}\left(\hat{a}^{(+)}\right) < \hat{S}\left(\hat{a}^{(-)}\right)$. При $s_{xy} < 0$ имеем $\hat{a}^{(+)} < 0, \hat{a}^{(-)} > 0$, а значит, как и в предыдущем случае, $\hat{S}\left(\hat{a}^{(+)}\right) < \hat{S}\left(\hat{a}^{(-)}\right)$.

Для полноты доказательства покажем теперь, что значения (10) являются точкой экстремума функции $\tilde{S}(a,b)=\dfrac{\sum_{i=1}^{n}(y_i-ax_i-b)^2}{\sigma_y^2+a^2\sigma_x^2}$, в которой она принимает наименьшее значение. Для этого, вводя обозначения

$$A=\left.\frac{\partial^2\tilde{S}}{\partial a^2}\right|_{\substack{a=\hat{a},\\ b=\hat{b}}},\ B=\left.\frac{\partial^2\tilde{S}}{\partial a\partial b}\right|_{\substack{a=\hat{a},\\ b=\hat{b}}},\ C=\left.\frac{\partial^2\tilde{S}}{\partial b^2}\right|_{\substack{a=\hat{a},\\ b=\hat{b}}},$$

достаточно показать справедливость неравенств

$$A>0,\ C>0,\ D=B^2-AC<0.$$

Выражения для производных суть

$$\frac{\partial^2\tilde{S}}{\partial a^2}=\frac{2}{\left(\sigma_y^2+a^2\sigma_x^2\right)}\left\{\sum_{i=1}^{n}x_i^2+\frac{4a\sigma_x^2\sum_{i=1}^{n}x_i\left(y_i-ax_i-b\right)}{\left(\sigma_y^2+a^2\sigma_x^2\right)}-\right.$$

$$\left.-\frac{\sigma_x^2\left(\sigma_y^2-3a^2\sigma_x^2\right)\sum_{i=1}^{n}\left(y_i-ax_i-b\right)^2}{\left(\sigma_y^2+a^2\sigma_x^2\right)^2}\right\},$$

$$\frac{\partial^2\tilde{S}}{\partial a\partial b}=\frac{2}{\left(\sigma_y^2+a^2\sigma_x^2\right)}\left\{\sum_{i=1}^{n}x_i+\frac{2a\sigma_x^2\sum_{i=1}^{n}\left(y_i-ax_i-b\right)}{\left(\sigma_y^2+a^2\sigma_x^2\right)}\right\},$$

$$\frac{\partial^2\tilde{S}}{\partial b^2}=\frac{2n}{\left(\sigma_y^2+a^2\sigma_x^2\right)}.$$

Тогда, учитывая, что $\hat{a}$, $\hat{b}$ (10) удовлетворяют (9”), получим

$$A=\frac{2n}{\left(\sigma_y^2+\hat{a}^2\sigma_x^2\right)}\left\{\frac{2\sigma_x^2 s_{xy}^2}{\left(\sigma_x^2 s_y^2-\sigma_y^2 s_x^2\right)+\sqrt{\left(\sigma_x^2 s_y^2-\sigma_y^2 s_x^2\right)^2+4\sigma_x^2\sigma_y^2 s_{xy}^2}}+\left(\overline{x}\right)^2\right\}>0,$$

$$B=\frac{2n\overline{x}}{\left(\sigma_y^2+\hat{a}^2\sigma_x^2\right)},$$

$$C=\frac{2n}{\left(\sigma_y^2+\hat{a}^2\sigma_x^2\right)}>0,$$

$$D=\frac{-8n^2\sigma_x^2 s_{xy}^2}{\left(\sigma_y^2+\hat{a}^2\sigma_x^2\right)^2\left\{\left(\sigma_x^2 s_y^2-\sigma_y^2 s_x^2\right)+\sqrt{\left(\sigma_x^2 s_y^2-\sigma_y^2 s_x^2\right)^2+4\sigma_x^2\sigma_y^2 s_{xy}^2}\right\}}<0.$$

Что и требовалось доказать. Отметим, что вид оценок (10) эквивалентен оценкам, получаемым по методу максимального правдоподобия при предположении нормальности ошибок [1, Гл.29].

Следствие 2. В случае $\sigma_x^2=\sigma_y^2$ задача (7) ((7')) имеет единственное решение, удовлетворяющее уравнениям

$$\hat{b}_{opm}=\overline{y}-\hat{a}_{opm}\overline{x},$$

$$\hat{a}_{opm}^2 s_{xy}-\hat{a}_{opm}\left(s_y^2-s_x^2\right)-s_{xy}=0,$$

и определяется равенствами

$$\hat{b}_{opm}=\overline{y}-\hat{a}_{opm}\overline{x},$$

$$\hat{a}_{opm}=\frac{\left(s_y^2-s_x^2\right)+\sqrt{\left(s_y^2-s_x^2\right)^2+4s_{xy}^2}}{2s_{xy}}= \tag{11}$$

$$=\frac{\left(\frac{s_y}{s_x}-\frac{s_x}{s_y}\right)+\sqrt{\left(\frac{s_y}{s_x}-\frac{s_x}{s_y}\right)^2+4r_{xy}^2}}{2r_{xy}}.$$

Величины $\hat{a}_{opm}$ и $\hat{b}_{opm}$ являются оценками коэффициентов a, b «ортогональной» регрессии [2].

Следствие 3. В случае $\sigma_x^2=0$ решением задачи (7) ((7')) являются оценки «прямой» регрессии

$$\hat{b}_{пр} = \overline{y} - \hat{a}_{пр}\overline{x},$$
$$\hat{a}_{пр} = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}. \qquad (12)$$

Следствие 4. В случае $\sigma_y^2 = 0$ решением задачи (7) ((7')) являются оценки «обратной» регрессии

$$\hat{b}_{обр} = \overline{y} - \hat{a}_{обр}\overline{x},$$
$$\hat{a}_{обр} = \frac{s_y^2}{s_{xy}} = \frac{1}{r_{xy}}\frac{s_y}{s_x}. \qquad (13)$$

Следствие 5. Оценки $\hat{a}$ и $\hat{b}$ (10) могут быть выражены через оценки «прямой» и «обратной» регрессий в виде

$$\hat{b} = \overline{y} - \hat{a}\overline{x},$$
$$\hat{a} = \frac{\left(s_y^2 - \frac{\sigma_y^2}{\sigma_x^2}s_x^2\right) + \sqrt{\left(s_y^2 - \frac{\sigma_y^2}{\sigma_x^2}s_x^2\right)^2 + 4\frac{\sigma_y^2}{\sigma_x^2}s_{xy}^2}}{2s_{xy}} = \qquad (14)$$
$$= \frac{\left(\hat{a}_{обр} - \frac{\sigma_y^2}{\sigma_x^2\hat{a}_{пр}}\right) + \sqrt{\left(\hat{a}_{обр} - \frac{\sigma_y^2}{\sigma_x^2\hat{a}_{пр}}\right)^2 + 4\frac{\sigma_y^2}{\sigma_x^2}}}{2}.$$

Следствие 6. Если в качестве приближений дисперсий ошибок σ_x^2, σ_y^2 взять их выборочные аналоги s_x^2, s_y^2, то оценки $\hat{a}$ и $\hat{b}$ (10) являются оценками «диагональной» регрессии [2]

$$\hat{b}_{диаг} = \overline{y} - \hat{a}_{диаг}\overline{x},$$
$$\hat{a}_{диаг} = \sqrt{\frac{s_y^2}{s_x^2}} = \sqrt{\hat{a}_{пр}\hat{a}_{обр}}. \qquad (15)$$

Таким образом, доказанное утверждение обобщает ранее известные оценки параметров линейной парной регрессии и дает новый явный вид МНК-оценок $\hat{a}$ и $\hat{b}$ (10) в случае наличия ошибок в обеих переменных X, Y.

Свойства оценок параметров. Остановимся лишь на некоторых свойствах оценок (10).

Утверждение 2. Справедливы соотношения

$$\left|\hat{a}_{пр}\right| \le \left|\hat{a}\right| \le \left|\hat{a}_{обр}\right|,$$

причем равенства имеют место лишь в случае точной линейной зависимости переменных X, Y.

Доказательство аналогично доказательству соотношения $\left|\hat{a}_{пр}\right| \le \left|\hat{a}_{орт}\right| \le \left|\hat{a}_{обр}\right|$ в [2].

Докажем теперь, аналогично тому как это сделано в [1, Гл.29], что оценки (10), как и (11) [2], являются состоятельными, в отличие от оценок (12), (13) в случае ошибок в обеих переменных X, Y $(\sigma_x^2 \neq 0,\ \sigma_y^2 \neq 0)$.

Утверждение 3. При условии $\operatorname{cov}\left(e_x, e_y\right) = 0$ и $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \xrightarrow[n\to\infty]{} \Sigma_X^2$ оценки $\hat{a}$ и $\hat{b}$ (10) являются состоятельными.

Доказательство. Используя (4') и свойства дисперсий и ковариаций, а также учитывая, что X, Y не являются случайными величинами, имеем следующие разложения

$$s_x^2 = s_{X+e_x}^2 = S_X^2 + s_{e_x}^2 + 2s_{X,e_x},$$
$$s_y^2 = s_{aX+b+e_y}^2 = a^2 S_X^2 + s_{e_y}^2 + 2as_{X,e_y},$$
$$s_{xy} = s_{(X+e_x),(aX+b+e_y)} = aS_X^2 + s_{e_x,e_y} + s_{X,e_y} + as_{X,e_x}.$$

Воспользуемся тем, что выборочные дисперсии и ковариация с ростом объема выборки сходятся по вероятности к своим математическим ожиданиям, то есть стремятся к своим теоретическим аналогам, если последние существуют [3, Гл.10]. В нашем случае, поскольку $s_{X,e_x} \to 0,\ s_{X,e_y} \to 0, n \to \infty$, имеем:

$$s_x^2 \to \Sigma_X^2 + \sigma_x^2,\ s_y^2 \to a^2\Sigma_X^2 + \sigma_y^2,\ s_{xy} \to a\Sigma_X^2,\ n \to \infty.$$

Теперь, подставляя в (10), имеем

$$\hat{a} \to \frac{\Sigma_X^2\left(a^2\sigma_x^2-\sigma_y^2\right)+\sqrt{\Sigma_X^4\left(a^2\sigma_x^2-\sigma_y^2\right)^2+4a^2\sigma_x^2\sigma_y^2\Sigma_X^4}}{2a^2\sigma_x^2\Sigma_X^2} =$$

$$= \frac{\left(a^2\sigma_x^2-\sigma_y^2\right)+\left(a^2\sigma_x^2+\sigma_y^2\right)}{2a^2\sigma_x^2} = a,\ n\to\infty.$$

Отсюда следует, что оценка $\hat{a}$ состоятельная.

Докажем теперь состоятельность оценки $\hat{b}$. Воспользуемся тем фактом, что $\overline{y} \to Y = aX + b,\ \overline{x} \to X, n \to \infty$ [3, Гл.10]. Тогда $\hat{b} = \overline{y} - \hat{a}\overline{x} \to aX + b - aX = b,\ n \to \infty$. Оценка $\hat{b}\,(10)$ - состоятельная.
Что и требовалось доказать.

Численные иллюстрации

Пример 1 *(дисперсии ошибок заданы).* Пусть истинные значения параметров модели (1) $a = 0.5, b = 4$ и $\left\{X_i, i = \overline{1,6}\right\} = \{6,7,8,9,10,11\}$. С помощью датчика случайных чисел нормального закона распределения получим выборку n значений ошибок $e_{xi} \sim N(0,1)$ и n значений $e_{yi} \sim N(0,3)$, а затем рассчитаем x_i, y_i по правилу (2), (3). Предполагая известными дисперсии ошибок $\sigma_x^2 = 1, \sigma_y^2 = 3$, для параметров a, b конфлюэнтной модели (4) по выборке $\left(x_i, y_i\right), i = \overline{1,n}$, были вычислены оценки (10), (12), (13), (15). При различных n полученные результаты приведены в табл.1.

Табл.1. Оценки параметров линейной модели при нормально распределенных ошибках.

Метод оценивания	n=300		n=30000	
	a=0.5	b=4	a=0.5	b=4
(10)	0.4985	4.0124	0.5	4.004
(12)	0.3694	5.1099	0.372	5.085
(13)	2.6019	-13.866	2.552	-13.439
(15)	0.9804	-0.0835	0.975	-0.035

Восстановленные по оценкам графики линейных приближений в случае $n = 300$ изображены на рис.1. «Крестиками» отмечены точки с истинными значениями $\left\{\left(X_i, Y_i\right), i = \overline{1,6}\right\}$, а «ромбиками» моделируемые точки с ошибками в обеих координатах $\left\{\left(x_i, y_i\right), i = \overline{1,6}\right\}$.

Как и ожидалось, наилучшее приближение к истинным значениям $a = 0.5, b = 4$ дают оценки (10).

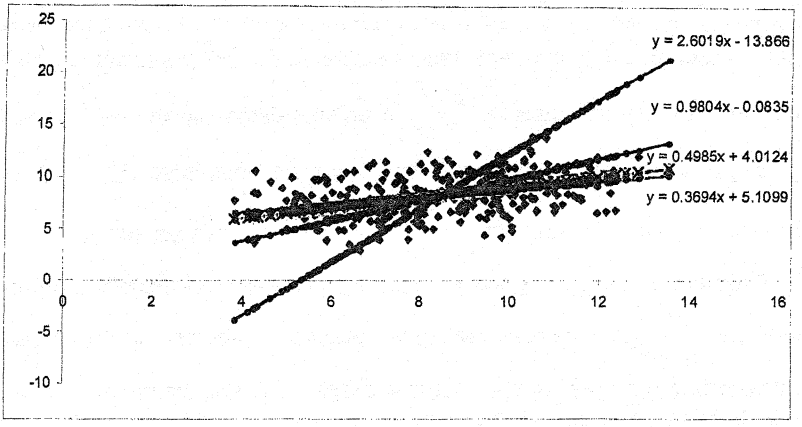

Рис.1

Пример 2 *(дисперсии ошибок заданы).* Аналогично первому примеру с помощью датчика случайных чисел получим выборку n значений равномерно распределенных ошибок e_{xi}, с $Ee_{xi} = 0, De_{xi} = 1$, и n значений ошибок e_{yi}, с $Ee_{yi} = 0, De_{yi} = 3$, а затем рассчитаем x_i, y_i по правилу (2), (3). Предполагая известными дисперсии ошибок $\sigma_x^2 = 1, \sigma_y^2 = 3$, для параметров a, b конфлюэнтной модели (4) по выборке $(x_i, y_i), i = \overline{1,n}$, были вычислены оценки (10), (12), (13), (15). При различных n результаты расчетов приведены в табл.2.

Табл. 2. Оценки параметров линейной модели при равномерно распределенных ошибках.

Метод оценивания	n=300		n=30000	
	a=0.5	b=4	a=0.5	b=4
(10)	0.502	3.980	0.498	4.014
(12)	0.373	5.083	0.371	5.093
(13)	2.582	-13.695	2.557	-13.487
(15)	0.981	-0.087	0.975	-0.034

Как и в предыдущем примере, наилучшее приближение к истинным значениям $a = 0.5, b = 4$ дают оценки (10).

Пример 3 *(дисперсии ошибок неизвестны, но известно их отношение).* На практике редко встречается ситуация когда дисперсии ошибок известны. Однако для применения $\hat{a}, \hat{b}$ (10) достаточно знать отношение дисперсий ошибок σ_y^2 / σ_x^2 и использовать равенство (10) или (14). В рассмотренных выше примерах вместо значений $\sigma_x^2 = 1, \sigma_y^2 = 3$ достаточно задать их отношение $\sigma_y^2 / \sigma_x^2 = 3$ и результат будет тот же.

Примером, где дисперсии ошибок неизвестны, а известно лишь их отношение σ_y^2 / σ_x^2, может служить задача оценивания линейной зависимости $Y = aX + b$ между ростом детей (Y) и средним ростом их родителей (X) [4]. При объеме выборки $n = 928$ данные группировались по 11 группам среднего роста у родителей и 14 группам роста у детей и приведены в табл.3. В верхней строке указан средний рост родителей в дюймах соответствующей группы, а по строкам – общее количество взрослых детей с указанным в левом столбце их росте в дюймах.

Табл.3. Данные о среднем росте родителей и их взрослых детей.

		Средний рост родителей (в дюймах)										
Рост взрослых детей (в дюймах)		63.5	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	73.5
	74.2							5	3	2	4	
	73.2						3	4	3	2	2	3
	72.2			1		4	4	11	4	9	7	1
	71.2			2		11	18	20	7	4	2	
	70.2			5	4	19	21	25	14	10	1	
	69.2	1	2	7	13	38	48	33	18	5	2	
	68.2	1		7	14	28	34	20	12	3	1	
	67.2	2	5	11	17	38	31	27	3	4		
	66.2	2	5	11	17	36	25	17	1	3		
	65.2	1	1	7	2	15	16	4	1	1		
	64.2	4	4	5	5	14	11	16				
	63.2	2	4	9	3	5	7	1	1			
	62.2		1		3	3						
	61.2	1	1	1			1		1			

Предполагая постоянной дисперсию роста взрослых людей и равной σ_y^2, будем иметь дисперсию среднего роста родителей, как дисперсию половины суммы роста людей, равной $\sigma_y^2/2 = \sigma_x^2$, а дисперсию роста детей равной σ_y^2. Данному предположению $\sigma_y^2/\sigma_x^2 = 2$ близко отношение выборочных дисперсий роста детей s_y^2 и среднего роста родителей s_x^2 равное $s_y^2/s_x^2 = 1.97$. Для оценки параметров линейной модели применим формулы (12), (13), (15) и (10) с $\sigma_y^2/\sigma_x^2 = 2$. Полученные результаты вычислений представлены в табл.4.

Табл.4. Оценки параметров
линейной модели для данных в табл.3.

Метод оценивания	$n = 928$	
(10)	1.3548	-24.424
(12)	0.6455	24.004
(13)	3.0488	-140.151
(15)	1.4029	-27.726

На рис.2 в виде точек изображены исходные данные из табл.3 и их предполагаемые погрешности. По оценкам параметром, указанным в табл.4, построены прямые, соответствующие приближенным линейным моделям. Оценки (15) «диагональной» регрессии близки по значению с оценками (10) и поэтому соответствующий им график приближенной линейной модели не изображен на рис.2.

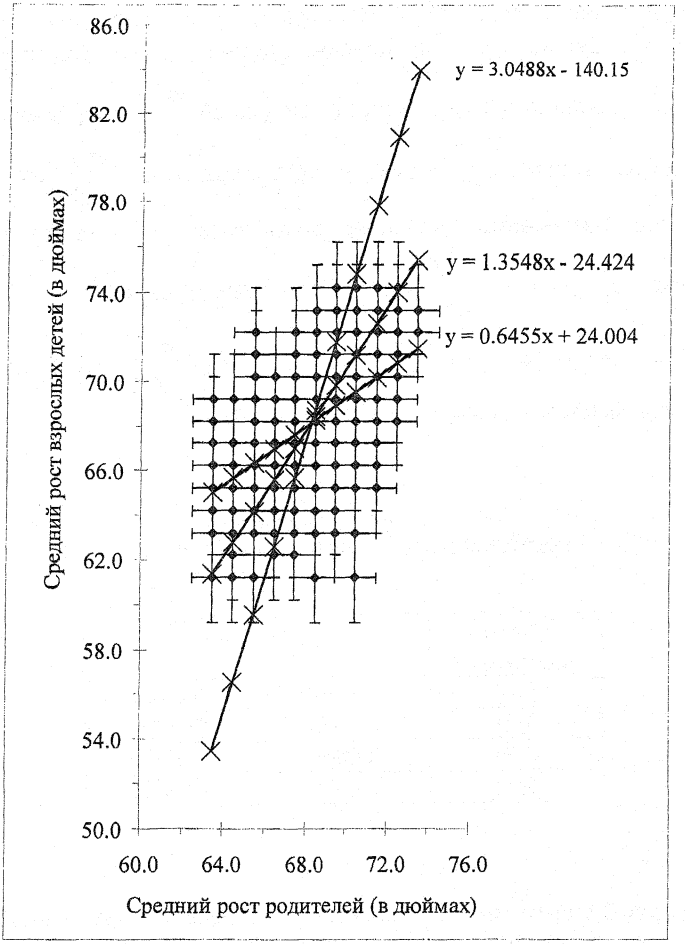

Рис.2.

Galton [4] по результатам своих исследований сделал вывод о «регрессивной» линейной зависимости роста детей от роста их родителей, что соответствует оценкам (12) «прямой» регрессионной модели. В то время как оценки (10) конфлюэнтной модели показывают «прогрессивную» линейную зависимость ростов.

Таким образом, полученные точечные МНК-оценки (10) для параметров простой линейной конфлюэнтной модели являются

наилучшими по сравнению с известными оценками «прямой», «обратной», «ортогональной», «диагональной» регрессии, а также эквивалентны оценкам, получаемым по методу максимального правдоподобия при предположении нормальности ошибок.

Литература

1. Кендалл М., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973.
2. Маленво Э. Статистические методы эконометрии. М.: Статистика, 1975.
3. Кендалл М., Стьюарт А. Теория распределений. М.: Наука, 1966.
4. Galton F. Regression towards mediocrity in hereditary stature // J. of the Royal Anthropolog. Inst., 1885. V.15. P.228-262.