УДК 519.25 DOI: 10.17587/it.26.555-563

В. Г. Зайцев, вед. инженер, e-mail: zaycev@looch.ru,
Научно-производственное предприятие геофизической аппаратуры "Луч", г. Новосибирск

К оцениванию параметров моделей при наличии ошибок во входной и выходной переменных. Линейный случай

Рассматриваются способы идентификации систем, описываемых моделями с погрешностями регистрации входной и выходной переменных. Считается, что при отсутствии информации о параметрах погрешностей в случае больших их значений получить приемлемые оценки всех искомых параметров модели без дополнительных, в сравнении с регрессионным анализом, предположений нельзя. Определение параметров регрессии при наличии погрешностей во входных и выходных переменных актуально при решении многих задач обработки данных. Во всех существующих подходах либо принимаются допущения относительно помех в виде, например, заданных соотношений между дисперсиями помех, либо используется метод последовательных приближений.

Для определения параметров линейной зависимости предложено использовать условие симметрии совместной плотности вероятности наблюдаемых входной и выходной переменных в косоугольных координатах.

Показана состоятельность и несмещенность оценки параметра связи переменных. Для случая нормальности помех в статье приведена формула определения параметра связи через оценки семиинвариантов четвертого порядка.

Ключевые слова и фразы: *структурный анализ, симметричное распределение помех, косоугольная система координат, состоятельность оценки*

Введение

Рассматриваются способы идентификации систем, описываемых моделями с погрешностями регистрации входной и выходной переменных. Если обе переменные системы случайны, то определение связей между ними обычно называют структурным анализом, а если входная переменная неслучайна — конфлюентным анализом. Считается, что при отсутствии информации о параметрах погрешностей в случае больших их значений получить приемлемые оценки всех искомых параметров модели без дополнительных, в сравнении с регрессионным анализом, предположений нельзя [1, 2]. Во всех существующих подходах либо принимаются допущения относительно помех в виде, например, заданных соотношений между дисперсиями помех, либо используется метод последовательных приближений.

Идентификация в одномерном линейном случае

Рассмотрим подробно возможные методы идентификации для одномерного случая. Пусть случайные величины (СВ) W и V связаны однозначной зависимостью $W = \Psi(V)$ и в пассивном эксперименте вместо W и V наблюдаются СВ

$$Y = W + E, \tag{1}$$
$$X = V + H, \tag{2}$$

где E и H — случайные независимые нормально распределенные ошибки измерений со свойствами $M[E] = 0$, $M[H] = 0$; $M[HE] = 0$, $M[HV] = 0$, $M[EV] = 0$, где M — символ математического ожидания.

Обозначим дисперсии СВ E, H и V соответственно $M[E]^2 = \sigma_\varepsilon^2$; $M[H]^2 = \sigma_\eta^2$; $M[V]^2 = \sigma_v^2$, а математическое ожидание СВ V обозначим $M[V] = m_v$.

Пусть в линейном случае структура системы задана соотношением

$$W = a + \theta V, \tag{3}$$

где a и θ — неизвестные коэффициенты, их значения необходимо оценить по выборке СВ X и Y объемом n. По наблюдаемым данным можно найти оценки параметров $R_{xy} = \theta\sigma_v^2$; $\sigma_x^2 = \sigma_v^2 + \sigma_\eta^2$; $\sigma_y^2 = \theta\sigma_v^2 + \sigma_\varepsilon^2$; $m_x = m_v$, $m_y = a + \theta m_v$, где R_{xy} — коэффициент ковариации; m_x, m_y — средние значения СВ X и Y. Модель не идентифицируется, поскольку для шести неизвестных a, θ, $\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_v^2$ и m_v имеем пять уравнений. Вычтем из СВ X и Y их средние значения и преобразуем СВ. В геометрической интерпретации преобразование представляет собой повороты осей Ox, Oy исходной системы координат (СК) к новым осям Ox_{01}, Oy_0:

$$X_{01} = [(X - m_x)\cos\beta + (Y - m_y)\sin\beta]\cos^{-1}(\alpha - \beta); \quad (4)$$

$$Y_0 = [-(X - m_x)\sin\alpha + (Y - m_y)\cos\alpha]\cos^{-1}(\alpha - \beta), \quad (5)$$

где α и β — углы поворота осей, причем $\mathrm{tg}\alpha = \theta$, а $\mathrm{tg}\beta = \mathrm{tg}\alpha\sigma_\eta^2\sigma_\varepsilon^{-2}$. Повороты оси Ox к оси Ox_0 на угол α и оси Oy к оси Oy_0 на угол β выполняются против хода движения часовой стрелки.

Подставив в соотношения (4) и (5) выражения (1)—(3), получим

$$X_{01} = [V - m_x + (a + \theta V - m_y)\mathrm{tg}\beta]\cos^{-1}\alpha\cos^{-1}(\alpha - \beta) + H_0;$$

$$Y_0 = [-(V - m_x)\mathrm{tg}\alpha + (a + \theta V - m_y)]\cos^{-1}\beta\cos^{-1}(\alpha - \beta) + E_0,$$

где $H_0 = (H\cos\beta + E\sin\beta)\cos^{-1}(\alpha - \beta)$ и $E_0 = (-H\sin\alpha + E\cos\alpha)\cos^{-1}(\alpha - \beta)$.

Проведя тригонометрические преобразования, с учетом равенства $\mathrm{tg}\alpha = \theta$ придем к выражениям

$$X_{01} = H_0 + V_0 + c_1;$$

$$Y_0 = E_0,$$

где $V_0 = V\cos^{-1}\alpha$, $c_1 = [(a - m_y)\mathrm{tg}\beta - m_x]\cos^{-1}\alpha \times \cos^{-1}(\alpha - \beta)$.

Перенесем параллельно себе ось Ox_{01} так, чтобы $X_0 = X_{01} - c_1$. Математическое ожидание произведения равно $M[H_0E_0] = M[(H\cos\beta + E\sin\beta)(-H\sin\alpha + E\cos\alpha)]\cos^{-2}(\alpha - \beta) = 0$. Сумма нормальных величин — величина нормальная, следовательно, является СВ. H_0 и E_0 независимы.

Рассмотрим плотность вероятности (ПВ) $f_{X_0Y_0}(x, y)$ системы СВ (X_0, Y_0). Эту плотность считаем симметричной относительно осей Ox_0, Oy_0, если

$$f_{X_0Y_0}(x, 0) = f_{X_0Y_0}(-x, 0) \quad (6)$$

$$\text{и } f_{X_0Y_0}(0, y) = f_{X_0Y_0}(0, -y). \quad (7)$$

В связи с независимостью СВ H_0 и E_0 ПВ $f_{H_0E_0}(\eta, \varepsilon)$ системы (H_0, E_0) будет симметрична по аналогии с соотношениями (6), (7).

Теорема 1. *Если плотность СВ V симметрична относительно математического ожидания, то при нормальной ПВ $f_{HE}(\eta, \varepsilon)$ существует единственная система координат с осями Ox_0, Oy_0, в которой ПВ $f_{X_0Y_0}(x, y)$ системы СВ (X_0,Y_0) симметрична относительно осей Ox_0 и Oy_0, одна из которых совпадает с линией $y = \theta x$. Если ПВ $f_V(v)$ несимметрична, то лишь в единственной указанной системе координат ПВ системы (X_0, Y_0) симметрична относительно оси, проходящей через линию $y = \theta x$.*

Во многих практических задачах регрессионного анализа при отсутствии погрешности независимой переменной условия применимости метода наименьших квадратов (МНК) нарушаются. Распространенным нарушением этих условий является неоднородность дисперсий наблюдаемой величины Y — гетероскедастичность данных. Встречаются ситуации, когда данные неоднородны по дисперсии, но их можно разделить на несколько групп однородных. Гетероскедастичность преодолевается с использованием взвешенного МНК (ВМНК) с многоэтапной процедурой оценивания коэффициентов регрессии на основе обычного МНК.

Несмотря на множество примеров использования ВМНК помеха E зависимой переменной математически в них не описывается: если она имеет различную дисперсию для различных значений независимой переменной, то формально она не случайная величина, а вектор. Однако это не мешает использовать ВМНК без формального определения этой помехи как вектора. Рассмотрим m неоднородных групп наблюдений, но с погрешностями в независимой и зависимой переменных.

Предположим, что СВ V в соотношениях (2) и (3) является дискретной величиной v_i, $i = 1, ..., m$, неважно, известной или нет. Число m этих значений может быть неизвестным. Свяжем с каждой парой v_i и $w_i = \theta v_i$ плотность вероятности $f_{iHE}(\eta, \varepsilon) = f_{iH}(\eta)f_{iE}(\varepsilon)$, при этом дисперсии нормально распределенных помех $\sigma_{i\eta}^2$ и $\sigma_{i\varepsilon}^2$ будем считать неизвестными, а их отношения $\sigma_{i\eta}^2/\sigma_{i\varepsilon}^2$ также неизвестным, но одинаковыми для всех i. Предельный случай подобного варианта: наличие m известных дискретных значений $v_i = i\Delta v$, где Δv — интервал между наблюдениями; $\sigma_{i\eta}^2 = 0$; $\sigma_{i\varepsilon}^2 = \sigma_\varepsilon^2 = \mathrm{const}$, $i = 1, ..., m$. Это представление наиболее часто используется в практике регрессионного анализа.

Следствие 1. *Теорема 1 выполняется в случае, если V представляет собой набор дискретных неизвестных значений v_i, $i = 1,...,m$, каждому из которых соответствует плотность ошибок $f_{iHE}(\eta,\varepsilon)$ с неизвестными в каждой точке значениями дисперсий $\sigma_{i\eta}^2$ и $\sigma_{i\varepsilon}^2$, но с одинаковым неизвестным отношением $\sigma_{i\eta}^2/\sigma_{i\varepsilon}^2$.*

Возможны различные алгоритмы получения оценок параметра θ с использованием свойств симметрии ПВ $f_{XY}(x, y)$. Эту оценку можно определить, сравнивая непосредственно эмпирическую плотность $f_{eXY}(x, y)$ по разную сторону осей СК.

Теорема 2. *Оценка параметра* θ, *определяемая по минимуму критерия Пирсона как критерия однородности выборок, относящихся к различным квадрантам координатной плоскости, в единственной системе координат, в которой ПВ* $f_{XY}(x, y)$ *симметрична, является состоятельной и асимптотически несмещенной.*

Доказательство теоремы 1. Доказательство проведем вначале для симметричной ПВ $f_v(v)$. До преобразования координат ПВ системы СВ (X, Y) при линейной связи W и V имеем

$$f_{XY}(x,y) = \int_{-\infty}^{\infty} f_V(v) f_H(x-v) f_E(y-\theta v)dv.$$

Поскольку в системе координат с осями Ox_0, Oy_0 СВ H_0 и E_0 также независимы, ПВ системы СВ (X_0, Y_0) равна

$$f_{X_0Y_0}(x_0,y_0) = f_E(y_0)\int_{-\infty}^{\infty} f_V(v_0) f_H(x_0-v_0)dv_0, \quad (8)$$

где $f_E(\varepsilon_0)$, $f_V(v_0)$, $f_H(\eta_0)$ находятся по правилам определения ПВ функций СВ.

ПВ $f_E(\varepsilon_0)$, $f_H(\eta_0)$ симметричны. Если ПВ $f_v(v_0)$ симметрична, то свертка (8) даст ПВ $f_{X_0Y_0}(x_0,y_0)$, зеркально симметричную относительно осей Ox_0, Oy_0.

Докажем теперь единственность угла $\alpha =$ $= \mathrm{arctg}\theta$, определяющего систему координат, в которой ПВ $f_{XY}(x, y)$ симметрична. Рассмотрим произвольную прямую $y = \theta_1 x$, $\theta_1 \neq \theta$ в системе координат с осями Ox_1, Oy_1. Преобразуем СВ X и Y по формулам (4), (5), только при этом вместо $\mathrm{tg}\alpha$ имеем $\mathrm{tg}\alpha_1 = \theta_1$, вместо $\mathrm{tg}\beta$ используем $\mathrm{tg}\beta_1$, а $V_1 = V/\cos\alpha_1$.

По аналогии с получением X_{01}, Y_0, H_0, E_0 преобразуем СВ X и Y:

$$X_{11} = V_1 c_2 + c_3 + H_1; \quad (9)$$

$$Y_{11} = V(\theta - \theta_1)\cos^{-1}(\alpha_1 - \beta_1) + \\ + m_x(\theta_1 - \theta)\cos^{-1}(\alpha_1 - \beta_1) + E_1, \quad (10)$$

где

$$c_2 = (\sigma_\varepsilon^2 + \theta\theta_1\sigma_\eta^2)/(\sigma_\varepsilon^2 + \theta_1^2\sigma_\eta^2);$$

$$c_3 = \{(1+\theta_1^2)[(a-m_y)\theta_1\sigma_\eta^2 - m_x\sigma_\varepsilon^2]\}/(\sigma_\varepsilon^2 + \theta_1^2\sigma_\eta^2).$$

Вычтем из X_{11} и Y_{11} их средние значения, обозначив разности через X_1 и Y_1.

Поскольку X_1 и Y_1 зависимы, то относительно осей Ox_1, Oy_1 ПВ $f_{X_1Y_1}(x_1,y_1)$ не будет симметричной.

Пусть теперь ПВ $f_v(v)$ несимметрична. Плотность $f_{xy}(x_0, y_0)$ определяется также по формуле (8). Так как $f_\varepsilon(y_0)$ симметрична, то ПВ $f_{xy}(x_0, y_0)$ симметрична относительно оси Ox_0, но несимметрична относительно оси Oy_0. Для угла поворота исходной СК $\alpha_1 \neq \alpha$ можно получить формулы, аналогичные формулам (9), (10) и соответствующие им выводы.

Доказательство следствия 1. В геометрической постановке в СК с осями Ox, Oy размещен набор ПВ $p_{i\eta\varepsilon}(x, y)$, $i = 1, ..., m$, с находящимися на линии $y = \theta x$ значениями математических ожиданий и отличающимися только ими от ПВ $f_{i\eta\varepsilon}(\eta, \varepsilon)$. Перейдем к косоугольной СК. Как и в предыдущем случае, положим $\mathrm{tg}\alpha = \theta$, а $\mathrm{tg}\beta = \mathrm{tg}\alpha\sigma_\eta^2\sigma_\varepsilon^{-2}$. Каждая из ПВ $p_{i\eta\varepsilon}(x, y)$ будет теперь расположена на оси Ox_0, и при этом математическое ожидание $M[E_{0i}] = 0$ и плотность симметрична относительно этой оси в данной СК. Таким образом, и в целом ПВ $f_{xy}(x, y)$, полученная усреднением всех ПВ $p_{i\eta\varepsilon}(x, y)$ в каждой точке косоугольной СК, будет в этой системе симметрична относительно оси Ox_0. Можно показать, что если значения v_i, $i = 1, ..., m$, заданы с одинаковым интервалом, и ПВ $f_{i\eta\varepsilon}(\eta, \varepsilon)$ одинаковы, то ПВ $f_{xy}(x, y)$ будет симметрична относительно обеих осей косоугольной СК.

Доказательство теоремы 2. Рассмотрим два положения осей СК: истинное — с осями Ox_0, Oy_0, тангенсами углов наклона относительно исходных осей Ox, Oy $\mathrm{tg}\alpha = \theta$, $\mathrm{tg}\beta = \mathrm{tg}\alpha\sigma_\eta^2\sigma_\varepsilon^{-2}$ и произвольное положение с отклонением осей Ox_1, Oy_1 от осей Ox_0, Oy_0 на углы $\Delta\alpha = \alpha_1 - \alpha$ и $\Delta\beta = \beta_1 - \beta$. Проведем одинаковое разбиение плоскостей этих СК на площадки, зеркально симметричные относительно осей, с числом площадок в каждом квадранте, равном r^2. Поскольку значения ПВ $f_{x_0y_0}(x_0,y_0)$ относительно осей системы координат с осями Ox_0, Oy_0 теоретически должны зеркально совпадать, то ее оценки в зеркально расположенных площадках можно считать принадлежащими выборкам из одной генеральной совокупности. Оценим однородность этих выборок. Для СК с осями Ox_0, Oy_0 статистика

$$\chi_{\lambda 0}^2 = \sum_{i=1}^{r}\sum_{j=1}^{r}\left[\frac{(k_{ij}^{(x_0,y_0)} - k_{ij}^{(x_0,-y_0)})^2}{k_{ij}^{(x_0,y_0)} + k_{ij}^{(x_0,-y_0)}} + \right. \\ + \frac{(k_{ij}^{(-x_0,y_0)} - k_{ij}^{(-x_0,-y_0)})^2}{k_{ij}^{(-x_0,y_0)} + k_{ij}^{(-x_0,-y_0)}} + \frac{(k_{ij}^{(y_0,x_0)} - k_{ij}^{(y_0,-x_0)})^2}{k_{ij}^{(y_0,x_0)} + k_{ij}^{(y_0,-x_0)}} + \quad (11) \\ \left. + \frac{(k_{ij}^{(-y_0,x_0)} - k_{ij}^{(-y_0,-x_0)})^2}{k_{ij}^{(-y_0,x_0)} + k_{ij}^{(-y_0,-x_0)}}\right]$$

асимптотически подчиняется χ^2-распределению с числом степеней свободы $q = 4r^2 - 1$. Здесь $k_{ij}^{(x_0,y_0)}$, $k_{ij}^{(x_0,-y_0)}$, $k_{ij}^{(-x_0,y_0)}$, $k_{ij}^{(-x_0,-y_0)}$, $k_{ij}^{(y_0,x_0)}$, $k_{ij}^{(y_0,-x_0)}$, $k_{ij}^{(-y_0,x_0)}$, $k_{ij}^{(-y_0,-x_0)}$ — числа попаданий в ij-ю площадку по обе стороны от совпадающей оси двух квадрантов; i, j — номера площадок вдоль осей, начиная от центра координат в сторону увеличения абсолютных значений координат.

Преобразуем первое слагаемое из соотношения (11) с учетом того, что для обоих квадрантов

$$k_{ij}^{(x_0,y_0)} = nsf_{e\,X_0Y_0}(x_{0\,ji}, y_{0\,ji}),$$

$$k_{ij}^{(x_0,-y_0)} = nsf_{e\,X_0Y_0}(x_{0\,ji}, -y_{0\,ji}),$$

где $f_{e\,X_0Y_0}(x_{0\,ji}, y_{0\,ji})$, $f_{e\,X_0Y_0}(x_{0\,ji}, -y_{0\,ji})$ — значения эмпирической плотности для ij-й площадки, s — площадь площадки:

$$\chi^2_{\lambda 01} =$$

$$= ns\sum_{i=1}^{r}\sum_{j=1}^{r}\frac{[f_{eX_0Y_0}(x_{0ij} - \Delta m_{x_0}, y_{0ij} - \Delta m_{y0}) -}{f_{eX_0Y_0}(x_{0ij} - \Delta m_{x_0}, y_{0ij} - \Delta m_{y0}) +} \to \quad (12)$$

$$\to \frac{-f_{eX_0Y_0}(x_{0ij} - \Delta m_{x0}, -y_{0ij} + \Delta m_{y0})]^2}{+f_{eX_0Y_0}(x_{0ij} - \Delta m_{x0}, -y_{0ij} + \Delta m_{y0})},$$

где $\Delta m_{x0} = (\Delta m_x \cos\beta_e + \Delta m_y \sin\beta_e)\cos^{-1}(\alpha_e - \beta_e)$; $\Delta m_{y0} = (-\Delta m_x \sin\alpha_e + \Delta m_y \cos\alpha_e)\cos^{-1}(\alpha_e - \beta_e)$, Δm_x, Δm_y — ошибки определения средних значений, α_e, β_e — углы поворота осей, соответствующие минимуму суммы значений выражений типа (12) для всех четырех сочетаний пар квадрантов.

Разность оценки плотности в симметрично взятых площадках квадрантов состоит из двух составляющих. Первая связана с ошибками оценивания плотности при $\Delta m_x = 0$ и $\Delta m_y = 0$, а вторая — со смещением нулей аргументов теоретической плотности. Ошибки определения средних значений СВ X и Y, а также X_0 и Y_0 с ростом n стремятся к нулю.

В СК с осями Ox_1, Oy_1 плотности $f_{eX_1Y_1}$ в соседних квадрантах отличаются, т. е. в среднем отличаются и числа попаданий в ij-ю площадку по обе стороны от совпадающей оси в формуле типа (11). При наличии средних значений в каждом слагаемом числителя эта статистика подчиняется нецентральному $\chi^2_{\lambda 1}$-распределению. В этом случае в формулу (12) вместо плотности $f_{eX_0Y_0}$ подставляется $f_{eX_1Y_1}$. Если вместо $f_{eX_1Y_1}$ использовать среднее значение на площадке теоретической плотности $f_{X_1Y_1}$, то получаемое при этом неслучайное значение статистики в сумме для четырех сочетаний пар квадрантов будет представлять собой параметр нецентральности $\lambda = 4ns\Delta f$, где Δf — значение суммы (12) при использовании теоретической ПВ $f_{X_1Y_1}$.

Для статистики $\chi^2_{\lambda 0}$ среднее значение и среднеквадратическое отклонение (СКО) равны $m_{\chi 0} = q$ и $\sigma_{\chi 0} = \sqrt{2q}$, а для $\chi^2_{\lambda 1}$ с нецентральным χ^2-распределением — $m_{\chi 1} = q + 4ns\Delta f$, $\sigma_{\chi 1} = \sqrt{2(q + 8ns\Delta f)}$.

Рассмотрим совместную ПВ $f_{\chi_1\chi_0}$ СВ $\chi^2_{\lambda 1}$ и $\chi^2_{\lambda 0}$. Вероятность превышения $\chi^2_{\lambda 0} > \chi^2_{\lambda 1}$ при $\Delta m_x = 0$ и $\Delta m_y = 0$ равна (аргументы x и y соответствуют $\chi^2_{\lambda 1}$ и $\chi^2_{\lambda 0}$ соответственно)

$$P = \int_{y=\chi^2_{\lambda 1}}^{\infty}\int_{x=0}^{\infty} f_{\chi_1\chi_0}(x, y)dxdy. \quad (13)$$

Интегрирование плотности проводится лишь выше линии $y = x$.

Рассмотрим квантили СВ $\chi^2_{\lambda 0}$ и $\chi^2_{\lambda 1}$: $K_{\lambda 0} = m_{\chi 0} + t\sigma_{\chi 0}$ и $K_{\lambda 1} = m_{\chi 1} - t\sigma_{\chi 1}$, где t — постоянная. Найдем их разность

$$\Delta K = 4ns\Delta f - t(\sqrt{2(q + 8ns\Delta f)} + \sqrt{2q}).$$

При $n \to \infty$ $\Delta m_x \to 0$ и $\Delta m_y \to 0$, а также при любом значении t разность $\Delta K \to \infty$. Эта разность характеризует расстояние между ПВ $f_{\chi 1}(x)$ и $f_{\chi 0}(x)$, чем оно больше, тем при меньших значениях пересекаются хвосты этих ПВ и меньше значение интеграла (13), так что при $n \to \infty$ вероятность $P \to 0$.

Таким образом, в случае определения положения осей по минимуму статистики χ^2_λ при $n \to \infty$ отклонения углов от истинных значений $|\Delta\alpha| \to 0$, $\Delta\beta \to 0$, а дисперсия оценки параметра θ также стремится к нулю, что является признаком состоятельности предлагаемой оценки.

Рассмотрим теперь смещенность оценок углов между осями Ox_0, Oy_0 и Ox, Oy, определяемых по минимуму статистики χ^2_λ.

В силу симметричности ПВ $f_{x_0y_0}(x_0, y_0)$ относительно осей, задаваемых углами α и β, совместное распределение ошибок оценивания средних значений Δm_{x0} и Δm_{y0} и ошибок оценки этой ПВ на противоположных относительно осей (оси для несимметричного распределения СВ V) ij-х площадках для бесконечного числа выборок конечного объема одинаковы. Следовательно, распределение всех случайных отклонений $\Delta\alpha$, $\Delta\beta$ будет симметричным относительно истинных значений, оценки углов α и β будут несмещенными. Оценка параметра θ равна $\theta_e = [\theta + \mathrm{tg}(\Delta\alpha)]/[1 - \theta\mathrm{tg}(\Delta\alpha)]$. При $n \to \infty$ $\mathrm{tg}(\Delta\alpha) \to 0$, т. е. оценка θ_e — асимптотически несмещенная.

Примечание. Теорему 1 со следствием и теорему 2 можно распространить и на случай коррелированных ошибок измерений.

Алгоритмы идентификации

Можно предложить несколько алгоритмов получения оценок параметра θ с использованием свойств симметрии ПВ $f_{xy}(x, y)$. Если оценку находить с использованием критерия Пирсона, то общий минимум определяется для суммы минимальных значений, рассчитанных по формуле (11) для всех четырех сочетаний пар квадрантов косоугольной СК, оси которой проведены через средние значения совокупности полученных точек. Оценка параметра a определяется после получения оценки θ_e:

$$a_e = m_{ye} - \theta_e m_{xe}. \quad (14)$$

При малом объеме выборки использование гистограммы в качестве оценки плотности затруднено. Эмпирическая ПВ СВ X и Y (гистограмма) представляется в виде ступенчатой функции. Если предполагать, что теоретическая ПВ имеет гладкую зависимость от аргументов x и y, то для уменьшения случайных ошибок эмпирической ПВ в виде ступенчатой функции целесообразно применить к ней процедуру сглаживания. При использовании популярного метода сглаживания с помощью ядерных функций возникает проблема выбора параметра локальности ядерной функции в зависимости от объема выборки. Эта проблема решается просто при использовании в качестве ядерных функций непрерывных ПВ без тяжелых хвостов [3]. Рассмотрим СВ

$$R = X + T; \quad (15)$$

$$Q = Y + U, \quad (16)$$

где T и U — независимые СВ с нормальными законами распределения, с нулевыми значениями математических ожиданий и заданными значениями дисперсий σ_t^2 и σ_u^2. Обозначим $f_{tu}(t, u)$ совместную плотность распределения СВ T и U. Эту плотность используем в качестве ядерной функции для массива значений выборки

$$f_{exy}(x, y) = n^{-1}\sum_{i=1}^{n} f_{tu}(x - x_i, y - y_i), \quad (17)$$

где x_i, y_i — значения выборки. Далее с учетом дисперсий σ_t^2 и σ_u^2 уменьшаем масштаб полученной оценки $f_{1xy}(x, y)$, получая сглаженную эмпирическую плотность выборки СВ X и Y с меньшей случайной ошибкой в сравнении со ступенчатой плотностью. Таким образом, за счет информации о гладкости ПВ уменьшается и случайная ошибка определения углов поворота осей.

Определим вид зависимости дисперсии оценки θ_e от характеристик СВ V, E, H и параметра θ. Симметричность ПВ $f_{xy}(x, y)$ в единственной СК означает симметричность в той же СК линий равной плотности вероятности. Рассмотрим произвольное сечение ПВ $f_{xy}(x, y$ и его характерные размеры вдоль осей Ox_0, Oy_0, они пропорциональны СКО СВ X_0 и Y_0. Дисперсии X_0 и Y_0 равны

$$\begin{aligned} D_{x0} &= M[(H\cos\beta + E\sin\beta + V_0)^2] = \\ &= (\theta^2+1)\sigma_v^2 + \sigma_\eta^2\sigma_\varepsilon^2(\sigma_\varepsilon^2 + \theta^2\sigma_\eta^2)/(\sigma_\varepsilon^4 + \theta^2\sigma_\eta^4), \end{aligned} \quad (18а)$$

$$\begin{aligned} D_{y0} &= M[(-H\sin\alpha + E\cos\alpha)^2] = \\ &= (\sigma_\varepsilon^2 + \theta^2\sigma_\eta^2)/(\theta^2+1). \end{aligned} \quad (18б)$$

Повернем ось Ox_0 на малые углы $\Delta\alpha$ и $-\Delta\alpha$ и обозначим полученные оси Ox_1 и Ox_2.

Уберем в каждой области горизонтального сечения ПВ $f_{xy}(x, y)$, прилегающей к оси Ox_1, симметричные к этой оси участки. Останутся центральносимметричные участки сечения (участки между осями Ox_1 и Ox_2). Отношение площади этих участков к площади сечения примерно пропорционально отношению СКО σ_{y0}/σ_{x0}. Эти отношения характеризуют рассеяние отклонений углов $\Delta\alpha$ и оценки параметра θ.

При малых углах отклонений найденной прямой $y = \theta_e x$ от истинного угла α дисперсию θ_e можно будет предполагать пропорциональной отношению дисперсий D_{y0}/D_{x0} из соотношений (18а) и (18б).

В качестве одного из используемых условий для первоначального оценивания углов поворота можно принять равенство нулю коэффициента ковариации СВ X_0 и Y_0. Это условие приводит к выражению для определения tgβ:

$$\mathrm{tg}\beta = (D_x\mathrm{tg}\alpha - R_{xy})/(D_y - R_{xy}\mathrm{tg}\alpha). \quad (19)$$

При этом для некоррелированных СВ E и H полагаем $0 \leqslant \beta \leqslant \pi/2$.

Таким образом, можно подбирать лишь угол α, а угол β будет вычисляться с использованием подбираемого угла α и оценок дисперсий и ковариации. Однако из-за наличия случайности в этих оценках и частного случайных величин можно предполагать несовпадение решений с независимым выбором углов и с использованием формулы (18), а также большее рассеяние результатов для последнего варианта.

Другие способы решения. Наличие информации о моментах погрешностей может упростить решение. Считаем СВ E и H нормальными. Для центрированных СВ $X_c = X - \bar{x}$ и $Y_c = Y - \bar{y}$, где $\bar{x}$ и $\bar{y}$ — математические ожидания, с учетом приведенных выше условий для СВ E и H и их нормальности определим вторые и четвертые моменты:

$$M(X_c^4) = M(V_c^4) + 6M(V_c^4 H^2) + M(H^4) =$$
$$= \mu_{4vc} + 6\sigma_\eta^2\sigma_v^2 + 6K_{v2\eta 2} + \mu_{4\eta},$$

где V_c — центрированная входная СВ; μ_{4vc} и $\mu_{4\eta}$ — центральные моменты четвертого порядка СВ V_c и H; $K_{v2\eta 2}$ – обозначение ковариации СВ V_c^2 и H^2. Функции независимых СВ независимы, поэтому $K_{v2\eta 2} = 0$. Выражение для четвертого момента примет вид

$$M(X_c^4) = \mu_{4vc} + 6\sigma_\eta^2\sigma_v^2 + 3\sigma_\eta^4.$$

Семиинвариант четвертого порядка СВ X_c равен

$$\rho_4(X_c) = M(X_c^4) - 3\sigma_\chi^4 = \mu_{4vc} - 3\sigma_v^4.$$

Аналогично определяется семиинвариант четвертого порядка СВ Y_c:

$$\rho_4(Y_c) = \theta^4[\mu_{4vc} - 3\sigma_v^4],$$

откуда для оценки параметра имеем

$$\theta_e = \sqrt[4]{\rho_{e4}(Y)/\rho_{e4}(X)}. \tag{20}$$

В работе [1] модифицированным методом наименьших квадратов для пассивной схемы экспериментов (ММНКП) оценки θ_e определяются с использованием последовательных приближений. На каждом этапе последовательных приближений в ММНКП используется обычный МНК.

Оценки параметров в ММНКП совпадают с оценками, полученными методом наименьших расстояний (МНР) [1, 2]. Утверждается (см., например, работу [2, стр. 244]), что в пассивном эксперименте при неизвестных значениях дисперсий помех построить состоятельные оценки параметра θ нельзя. Оценим в нестрогой постановке ошибки определения оценки θ_e с помощью МНР.

Для одномерной линейной зависимости типа (2) с дискретно заданной величиной v при нахождении оценки коэффициента θ при заданных значениях дисперсий помех в ММНКП с использованием выражения

$$q_e = \operatorname{Arg\,min}\left(n^{-1}\sum_{i=1}^{n}\frac{(y_i - \theta x_i)^2}{\sigma_\varepsilon^2 + \theta^2\sigma_\eta^2} \right) \tag{21}$$

дисперсию θ_e можно оценить в виде [2]

$$D_{e\theta} = \frac{(\sigma_\varepsilon^2 + \theta^2\sigma_\eta^2)}{\sum_{i=1}^{n}(x_i - m_{xe})^2}. \tag{22}$$

При некотором объеме выборки больший вклад в ошибку определения параметра вносит не зависящее от n смещение оценки. Считая плотность $f_v(v)$ симметричной, рассмотрим ПВ $f_{xy}(x, y)$ в такой прямоугольной СК с осями Ox_1 и Oy_1, в которой Ox_1 совпадает с прямой $y = \theta x$. ПВ $f_{xy}(x,y)$ получается суммированием бесконечного числа взятых с весом $f_v(v)dv$ плотностей $f_{\eta\varepsilon}(x, y)$ с математическими ожиданиями m_x и $m_y = \theta v$. Если СВ E и H некоррелированы, то при $\sigma_\varepsilon^2 \neq \sigma_\eta^2$ оси симметрии каждого горизонтального сечения ПВ $f_{\eta\varepsilon}(x, y)$ не совпадают с осями Ox_1 и Oy_1. Поэтому оси симметрии горизонтальных сечений плотности $f_{xy}(x, y)$ также не совпадают с этими осями. Уберем в каждой области горизонтального сечения ПВ $f_{xy}(x, y)$, прилегающей к оси Ox_1, симметричные к этой оси участки. Из-за наличия оставшихся после этого по обе стороны от оси центральносимметричных участков сечения можно сделать вывод, что определяемая для каждого сечения прямая $y = \theta_1 x$ с использованием МНР не совпадает с осью Ox_1. Соответственно и для ПВ $f_{x_1 y_1}(x, y)$ в данной прямоугольной СК параметр θ_1 не равен нулю, и оценка θ_e по МНР смещена, кроме случаев, когда $\sigma_\varepsilon^2 = \sigma_\eta^2$ или когда одна из осей симметрии горизонтальных сечений ПВ $f_{\eta\varepsilon}(x, y)$ перпендикулярна к линии $y = \theta x$ при коррелированности СВ E и H.

В работе [4] предложен способ, основанный на знании коэффициента эксцесса погрешности η входной переменной. Для определения коэффициента θ этот метод в одномерном случае в предположении нормальности СВ η можно свести к использованию формулы

$$\theta_e = \theta_{11}(\theta_{11}\theta_{12}\mu_{4ex} - 3R_{exy}^2)/(\theta_{11}\mu_{4ex} - 3R_{exy}^2), \tag{23}$$

где μ_{4ex} — оценка четвертого центрального момента СВ X; θ_{11} — МНК-оценка коэффициента θ; θ_{12} — оценка коэффициента θ, полученная взвешенным МНК.

В работе [5] предложена оценка с использованием смешанных моментов:

$$\theta_e = M[YYX]/M[YXX]. \tag{24}$$

Недостаток двух последних подходов состоит в большей относительно МНР дисперсии оценки θ_e, в основном обусловленной наличием частного оценок.

В качестве примера рассмотрим последнюю формулу. Распишем ее для конечного объема выборки:

$$\theta_e = \frac{\sum_{i=1}^{n} (\theta v_i + \varepsilon_i)^2 (v_i + \eta_i)}{\sum_{i=1}^{n} (\theta v_i + \varepsilon_i)(v_i + \eta_i)^2}.$$

С использованием линеаризации функции случайных величин $Z = \varphi(Z_1, Z_2)$ для среднего значения и дисперсии получим

$$\bar{z} = \bar{z}_1 / \bar{z}_2;$$

$$\sigma_z^2 = \sigma_{z2}^2 / \bar{z}_2^2 -$$

$$- (\bar{z}_1 / \bar{z}_2^2)^2 \sigma_{z1}^2 - 2(\bar{z}_1 / \bar{z}_2^3) K_{z1z2}.$$

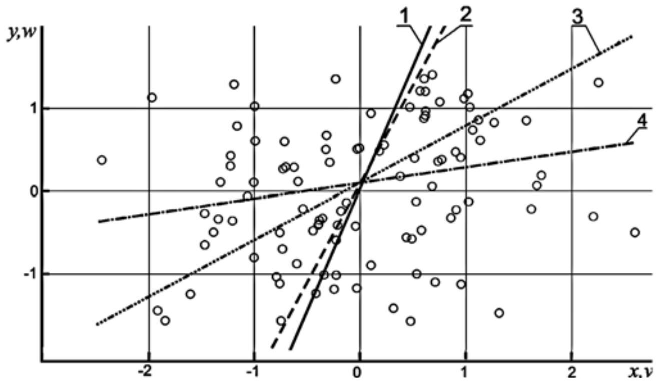
При равном нулю математическом ожидании СВ V оценка θ_e теоретически имеет распределение, для которого среднее значение и дисперсия не существуют. На практике это означает возможность получения больших выбросов этой оценки при малых значениях математических ожиданий СВ V.

Таблица 1

Расчет параметра связи при неслучайной входной величине

	Использование симметрии ПВ		МНР	$\theta_e = \sqrt[4]{\rho_{e4}(Y)/\rho_{e4}(X)}$	$\theta_e = M[YYX]/M[YXX]$	По формуле (23)
	С формулой (19)	Совместный выбор углов				
$\sigma_\varepsilon = 0{,}173$, $\sigma_\eta = 0{,}173$						
$\bar{\theta}$	1,0024	1,0075	1,0087	1,0101	0,1961	0,9670
σ_θ	0,0747	0,0520	0,0513	0,0756	1,3120	0,1328
m_{e2}	0,0056	0,0028	0,0027	0,0058	2,3611	0,0187
$\sigma_\varepsilon = 0$, $\sigma_\eta = 0{,}173$						
$\bar{\theta}$	0,9824	1, 0049	0,9578	1,0004	0,5102	1,0103
σ_θ	0, 0493	0,0285	0,0401	0,0574	2,7890	0,1012
m_{e2}	0,0027	0,0008	0,0034	0,0033	8,0285	0,0102
$\sigma_\varepsilon = 0{,}173$, $\sigma_\eta = 0$						
$\bar{\theta}$	1,0226	1,0077	1,0444	1,0139	1,5850	0,9980
σ_θ	0,0480	0,0291	0,0342	0,0464	1,1511	0,0935
m_{e2}	0,0028	0,0009	0,0031	0,0024	1,6641	0,0087
$\sigma_\varepsilon = 0$, $\sigma_\eta = 0{,}5$						
$\bar{\theta}$	0,9531	0,9902	0,7146	1,0313	0,6911	1,0364
σ_θ	0,1323	0,0911	0,0745	0,3135	2,6338	2,7354
m_{e2}	0,0197	0,0083	0,0870	0,0993	7,0326	7,4839

Примеры. 1. Входную величину v брали с интервалом 0,04 в диапазоне от −1,0 до 1,0. Были промоделированы нормально распределенные помехи с объемом выборки n = 51. Дисперсии погрешностей принимали равными $\sigma_\varepsilon^2 = 0{,}03$ и $\sigma_\eta^2 = 0{,}03$; $\sigma_\varepsilon^2 = 0$ и $\sigma_\eta^2 = 0{,}03$; $\sigma_\varepsilon^2 = 0{,}03$ и $\sigma_\eta^2 = 0$; $\sigma_\varepsilon^2 = 0$ и $\sigma_\eta^2 = 0{,}25$. Искомые параметры принимали равными a = 0 и θ = 1. В табл. 1 представлены результаты расчетов методом ММНКП/МНР, методом оценивания параметра θ, основанном на теоретическом отношении $\theta = M(YYX)/M(YXX)$, методом из работы [4] по формуле (23) и предложенными методами. При этом оси под углами α и β относительно старых осей проходят через точку, соответствующую средним значениям выборки СВ X и Y. В каждую точку (x_0, y_0) косоугольной системы координат, получаемую перебором углов поворота обеих осей, помещалась двумерная плотность нормального распределения с дисперсией по каждой координате $\sigma^2 = (\sigma_x^2 + \sigma_y^2) n^{0,66}$ и числом интервалов по каждой координате, равным 40. Для нахождения минимального значения критерия Пирсона в каждом положении осей суммировались квадраты разностей сглаженной ядерным способом оценки $f_{exy}(x_0, y_0)$ в четырех парах соседних квадрантов. Число значений выборок было равно 100. Для различных значений СКО ошибок в табл. 1 и в табл. 2 помещены среднее значение $\bar{\theta}$, СКО σ_θ и второй центральный момент m_{e2} оценки. Если СКО ошибок принимались одинаковыми, предлагаемый метод с перебором углов поворота обеих осей давал результа-



Истинная прямая (*1*); прямая, полученная предлагаемым методом (*2*); прямая, полученная МНР (*3*); прямая, полученная МНК (*4*)

ты, близкие к получаемым МНР. Однако при $\sigma_\varepsilon^2 \neq \sigma_\eta^2$ МНР ему существенно уступает.

Для МНР вначале использовали алгоритм из работы [2], а после этого, для увеличения точности оценки, в окрестности найденного значения проводили перебор значений θ с поиском минимальной суммы абсолютных значений отклонений исходных точек от прямой линии $y = a + \theta x$.

Средние значения оценки параметра a по формуле (14) с использованием предложенных в статье способов, а также МНР дали значения 0,001...0,004, а СКО оценок 0,02...0,08.

2. Приведем пример для случая, когда v распределена по равномерному закону в диапазоне от −1,0 до 1,0. На рисунке для $\theta = 3$, $n = 100$ и при $\sigma_\varepsilon = 0$, $\sigma_\eta = 1$ приведены типичные решения при использовании МНК, МНР и предложенного способа. Там же приведены истинная прямая и точки выборки.

3. Приведем пример для случая, когда v является СВ с несимметричным распределением.

Таблица 2

Расчет параметра связи при случайной входной величине, распределенной по логнормальному закону

Параметры ошибок	Параметры оценки	Использование симметрии ПВ: С использованием формулы (19)	Использование симметрии ПВ: Совместный выбор углов	МНР	Вычисление по формуле (23)	Вычисление по формуле (20)	Вычисление по формуле (24)
Объем выборки n = 500							
$\sigma_\varepsilon = 1{,}2$, $\sigma_\eta = 0{,}3$	$\bar{\theta}$	1,0451	1,0484	1,3239	1,0390	1,0381	1,0497
	σ_θ	0,0721	0,0489	0,0565	0,1016	0,1069	0,0596
	m_{e2}	0,0052	0,0024	0,0782	0,0104	0,0116	0,0036
$\sigma_\varepsilon = 0{,}4$, $\sigma_\eta = 1{,}3$	$\bar{\theta}$	1,0423	1,0470	0,8001	0,9648	1,0678	1,0501
	σ_θ	0,0663	0,0533	0,0347	0,1171	0,1344	0,0688
	m_{e2}	0,0045	0,0028	0,0638	0,0210	0,0184	0,0047
Объем выборки n = 50							
$\sigma_\varepsilon = 1{,}2$, $\sigma_\eta = 0{,}3$	$\bar{\theta}$	1,1481	1,1261	1,3266	1,1735	1,1421	1,0076
	σ_θ	0,2494	0,1298	0,1413	0,7650	0,4424	0,3451
	m_{e2}	0,0723	0,0227	0,0963	0,6010	0,2041	0,9210
$\sigma_\varepsilon = 0{,}4$, $\sigma_\eta = 1{,}3$	$\bar{\theta}$	0,9715	1,0030	0,8307	1,2061	1,0911	1,0122
	σ_θ	0,1871	0,1449	0,1328	1,8382	0,4910	1,7837
	m_{e2}	0,0358	0,0232	0,0657	3,3990	0,2435	3,1828

В работе [4] в модельном примере использовался логнормальный закон. Для сравнительной оценки методов будем также использовать его. Был проведен расчет для нормальных СВ ε и η с СКО, соответственно, 1,2 и 0,3; 0,4 и 1,3 и СВ v, распределенной по логнормальному закону LN(1; 0,5). При этом брались значения объема выборки n = 500 и n = 50, коэффициент θ = 1,05. Логнормальный закон относится к распределениям с тяжелыми хвостами, поэтому число экспериментов для n = 50 брали равным 500 для всех способов вычислений, кроме использования симметрии плотностей с независимым выбором углов, в этом случае число повторений было равно 100. В отличие от предыдущего примера здесь должны сравниваться характеристики сглаженной ядерным способом оценки $f_{exy}(x_0, y_0)$ лишь относительно оси Ox_0. Значения дисперсии по обеим координатам плотности нормального распределения, используемой в качестве ядра, принимались равными 2,25. Результаты расчетов сведены в табл. 2.

Из табл. 2 видно, что в предлагаемом способе с использованием симметрии плотности значение второго момента ошибки в основном определяется дисперсией оценки, а в МНР — смещением оценки.

Заключение

Метод с использованием симметричности плотностей гауссовых помех имеет превосходство над известными методами по точности. Возможно распространение предложенной методики на многомерный случай.

Список литературы

1. **Жилинская Е. И., Товмаченко Н. Н., Федоров В. В.** Методы регрессионного анализа при наличии ошибок в предикторных переменных. М.: АН СССР, Научный совет по комплексной программе "Кибернетика", 1979. С. 16—25.

2. **Айвазян С. А., Енюков И. С., Мешалкин Л. Д.** Прикладная статистика: Исследование зависимостей. М.: Финансы и статистика, 1985. С. 235—244.

3. **Зайцев В. Г.** К использованию ядерных оценок при сглаживании данных // Заводская лаборатория. Диагностика материалов. 2017. Т. 83, № 5. С. 66—71.

4. **Тимашев С. А., Тырсин А. Н.** Оценивание линейных структурных соотношений между случайными величинами // Заводская лаборатория. Диагностика материалов. 2010. Т. 76, № 3. С. 68—71.

5. **Gillard J. W.** Method of moments estimation in linear regression with errors in both variables // Communications in Statistics: Theory and Methods. 2014. Vol.43, N. 15. P. 3208—3222. URL: http://orca.cf.ac.uk/71432/.

V. G. Zaycev, Lead Engineer, e-mail: zaycev@looch.ru,
Scientific Production Enterprise of Geophysical Equipment "Looch"
Novosibirsk, 630051, Russian Federation

To Evaluation of Parameters of Models at Presence of Errors in Entrance and Output Variables. Linear Case

This paper presents the ways to identify systems described by models with errors in recording of input and output variables. It is considered that in the absence of information on the error parameters in the case of their large values, it is impossible to obtain acceptable estimates of all the required parameters of the model without additional, in comparison with the regression analysis, assumptions. The determination of the regression parameters in the presence of errors in the input and output variables is important for solving many tasks related to data processing. In all existing approaches, either the assumptions regarding the noise in the form of e.g. given relations between the noises dispersions are made, or the method of successive approximations is applied. To determine the parameters of the linear function, it is proposed to use the condition of the symmetry of the joint probability density of the observed input and output variables in the oblique coordinates. For the case of normality of the noise, the article gives a formula for determining the parameter of the relationship via the estimates of the fourth-order semi-invariants.

Keywords: *structural analysis, symmetric distribution of hindrances, oblique-angled system of coordinates, solvency of estimation*

DOI: 10.17587/it.26.555-563

References

1. **Zhilinskaya E. I., Tovmachenko I. S., Feodorov V. V.** Methods for regression analysis with errors in the predictor variables, Moscow, Nauka Publishers, 1979, 34 p. (in Russian).
2. **Aivazyan S. A., Yenukov I. S., Mechalkin L. D.** Applied statistics: Research of dependences, Moscow, Finance And Statistics Publishers, 1985, 487 p. (in Russian).
3. **Zaycev V. G.** On the Use of Kernel Estimators in Data-Smoothing, *Zavodskaya laboratoriya. Diagnostica materialov*, 2017, vol. 83, no. 5, pp. 66—71 (in Russian).
4. **Timachov S. A., Tyrsin A. N.** Estimation of linear structural relationships between the random variables, *Zavodskaya laboratoriya. Diagnostika materialov*, 2010, vol. 76, no. 3, pp. 68—71 (in Russian).
5. **Gillard J. W.** Method of moments estimation in linear regression with errors in both variables. *Communications in Statistics: Theory and Method,* 2014, vol. 43, no. 15, pp. 3208—3222, available at: http://orca.cf.ac.uk/71432/