УДК 517. 518.45

ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ КОРРЕЛЯЦИОННОЙ ЗАВИСИМОСТИ

д-р техн. наук, проф. С. Г. ЕХИЛЕВСКИЙ,
канд. физ.-мат. наук, доц. О. В. ГОЛУБЕВА, О. Н. ЗАБЕЛЕНДИК
(Полоцкий государственный университет, Беларусь)

Аннотация. *Методами теории вероятностей обоснована лаконичная процедура, позволяющая выразить параметры полиномиальной регрессии условного математического ожидания через смешанные статистические моменты системы случайных величин. Реализованы примеры линейной и квадратичной регрессии. Во втором случае рассмотрение ограничено ситуацией, когда плотность вероятности случайного аргумента является четной функцией. Результат получен без громоздких выкладок, ибо при его получении использованы не начальные статистические моменты, возникающие в методе наименьших квадратов, а смешанные центральные моменты, отражающие вид регрессионной кривой. Показано, что, в общем случае, учет нелинейности корреляционной зависимости лишь усиливает неравенство, подтверждающее адекватность регрессионного приближения. Обоснована сходимость такой процедуры, если условное математическое ожидание не является полиномом по сути.*

Ключевые слова: *условное математическое ожидание, полиномиальная регрессия, смешанные статистические моменты.*

Введение. Обычно регрессию корреляционной зависимости случайных величин осуществляют методом наименьших квадратов [1], не имеющим никакого отношения к теории вероятностей. К тому же его использование в случае нелинейной регрессии приводит к системам линейных алгебраических уравнений высоких порядков и как следствие громоздким выкладкам. По этой причине получение регрессионных кривых, как правило, просто не рассматривается [2]. Вместе с тем известно [3], что закон распределения случайной величины (или их системы) можно восстановить, пользуясь соответствующими статистическими моментами. В частности такая процедура реализована в [4] при моделировании рабочего процесса респиратора на химически связанном кислороде. Это позволило не только выделить асимптотику процесса, но и определить поправки к ней, обусловленные асимметриями и эксцессами высших порядков. Аналогично в [5] метод статистических моментов позволил рассмотреть независимые повторные испытания как асимптотически гауссовский случайный процесс с дискретным временем.

В настоящей публикации методами теории вероятностей обосновывается процедура, позволяющая выразить параметры линии регрессии условного матожидания через статистические моменты системы случайных величин. Рассматриваются случаи линейной и квадратичной регрессии.

Полиномиальная регрессия корреляционной зависимости. Рассмотрим двумерную случайную величину $\{X,Y\}$ с возможными значениями (x,y). Аппроксимируем условное математическое ожидание Y прямой линией

$$M(Y\,|\,X=x)=m_Y(x)\approx kx+b. \tag{1}$$

Ее параметры k и b выражаются через начальные и центральные моменты $\{X,Y\}$ независимо от их числовых значений.

$$b=m_Y-k\,m_X. \tag{2}$$

Подставив (2) в (1), сведем построение регрессионной прямой к отысканию ее углового коэффициента

$$m_Y(x)-m_Y\approx k(x-m_X). \tag{3}$$

По определению

$$m_Y(x)=\int_{-\infty}^{\infty} y\,f(y\,|\,x)\,dy, \tag{4}$$

где $f(y\,|\,x)$ – условная плотность вероятности Y. Умножим условие ее нормировки

$$1=\int_{-\infty}^{\infty} f(y\,|\,x)\,dy$$

на m_Y и отнимем полученное от (4):

$$m_Y(x)-m_Y=\int_{-\infty}^{\infty}(y-m_Y)f(y\,|\,x)\,dy=\frac{1}{f(x)}\int_{-\infty}^{\infty}(y-m_Y)f(x,y)\,dy. \tag{5}$$

Подставив (5) в (3), умножим результат на $(x-m_X)f(x)$ и выполним интегрирование по x:

$$\mu_{XY}\equiv\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\big(x-m_X\big)\big(y-m_Y\big)f(x,y)\,dx\,dy=k\int_{-\infty}^{\infty}(x-m_X)^2 f(x)\,dx=k\,\sigma_X^2 \tag{6}$$

Отсутствие после интегрирования зависимости от x позволяет подбором k обеспечить строгое равенство в (6). Именно такой выбор обеспечивает минимальность среднего значения условной дисперсии Y, полученной в приближении линейной

регрессии корреляционной зависимости, в чем можно убедиться[1], добавив к k произвольную поправку Δ

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(y-m_Y-(k+\Delta)(x-m_X)\right)^2 f(x,y)\,dx\,dy = \underbrace{\int_{-\infty}^{\infty}\left(y-m_Y\right)^2 f(y)\,dy}_{=\sigma_Y^2}\underbrace{\int_{-\infty}^{\infty} f(x|y)\,dx}_{=1} -$$

$$-2(k+\Delta)\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x-m_X\right)\left(y-m_Y\right)f(x,y)\,dx\,dy}_{=k\sigma_X^2\ (см.\,(6))} +$$

$$+(k+\Delta)^2\underbrace{\int_{-\infty}^{\infty}\left(x-m_X\right)^2 f(x)\,dx}_{=\sigma_X^2}\underbrace{\int_{-\infty}^{\infty} f(y|x)\,dy}_{=1} = \sigma_Y^2-2(k+\Delta)k\sigma_X^2+(k+\Delta)^2\sigma_X^2 =$$

$$=\sigma_Y^2-k^2\sigma_X^2+\Delta^2\sigma_X^2>\sigma_Y^2-k^2\sigma_X^2 \qquad (7)$$

Данный результат допускает следующую интерпретацию. Правильный выбор значений k и b обеспечивает совпадение точного значения смешанного центрального момента 2-го порядка μ_{XY} с полученным в приближении линейной регрессии корреляционной зависимости. Именно это обстоятельство исключает вклад в среднее значение условной дисперсии Y, обусловленный погрешностью, вносимой в μ_{XY} процедурой регрессии. При этом оставшаяся погрешность, связанная с процедурой регрессии обусловлена только смешанными центральными моментами более высоких порядков. Чтобы избавиться от нее нужно добавлять в (1) новые параметры (повышать степень регрессионного полинома).

С учетом соотношения (6), для углового коэффициента k, уравнение регрессии (3) примет вид

$$(m_Y(x)-m_Y)/\sigma_Y \approx r_{XY}(x-m_X)/\sigma_X\,, \qquad (8)$$

где

$$r_{XY}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{x-m_X}{\sigma_X}\cdot\frac{y-m_Y}{\sigma_Y}f(x,y)\,dx\,dy=\frac{\mu_{XY}}{\sigma_X\sigma_Y} \qquad (9)$$

– коэффициент линейной корреляции (центральный смешанный момент 2-го порядка приведенных случайных величин X и Y).

[1] Именно это является обоснованием того, что параметры k и b можно определять методом наименьших квадратов на основе экспериментальных данных по измерению двумерной случайной величины $\{X,Y\}$.

Согласно (6), (7) и (9) количественной характеристикой разброса $\{X,Y\}$, вокруг регрессионной прямой (8) является интеграл

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(y-m_Y-k(x-m_X)\right)^2 f(x,y)\,dx\,dy=\sigma_Y^2\left(1-r_{XY}^2\right). \tag{10}$$

Благодаря тому, что разброс $\{X,Y\}$ вокруг условного математического ожидания меньше, чем вокруг регрессионной прямой

$$M\left(\sigma_Y^2(x)\right)=\sigma_Y^2\left(1-D\left(m_Y(x)\right)\big/\sigma_Y^2\right)\le\sigma_Y^2\left(1-r_{XY}^2\right).$$

Иными словами,

$$\eta^2=D\left(m_Y(x)\right)\big/\sigma_Y^2\ge r_{XY}^2, \tag{11}$$

причем равенство возможно только в ситуации точного равенства (8), когда корреляция линейна по сути. В остальных случаях адекватность приближения (8) подтверждается силой неравенства

$$\left(\eta^2-r_{XY}^2\right)\big/\eta^2\ll 1. \tag{12}$$

Определенную согласно (11) величину η называют корреляционным отношением.

Если условие (12) нарушено, регрессия должна быть нелинейной, чтобы описывающий ее полином лучше «вписывался» в график условного матожидания $m_Y(x)$.

Тогда квадратичная регрессия корреляционной зависимости будет выглядеть так:

$$\frac{m_Y(x)-m_Y}{\sigma_Y}\approx r_{X^2Y}\,\frac{x^2-m_{X^2}}{\sigma_{X^2}}+r_{XY}\,\frac{x}{\sigma_X}. \tag{13}$$

Приближенное равенство (13) рассмотрено для случая, когда плотность вероятности X является четной функцией ($f(-x)=f(x)$) и получено без громоздких выкладок, ибо мы отталкивались не от начальных статистических моментов, возникающих в методе наименьших квадратов, а от смешанных центральных моментов, отражающих вид регрессионной кривой.

Покажем, что нелинейность регрессии снижает среднеквадратический разброс Y вокруг ее графика по сравнению с полученным в линейном приближении (см. (10):

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(y-m_Y-a\left(x^2-m_{X^2}\right)-b\left(x-m_X\right)\right)^2 f(x,y)\,dx\,dy=$$

$$= \underbrace{\int_{-\infty}^{\infty} \left(y - m_Y\right)^2 f(y)dy}_{=\sigma_Y^2} \underbrace{\int_{-\infty}^{\infty} f(x|y)dx}_{=1} -$$

$$-2 \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(y - m_Y\right)\left(a\left(x^2 - m_{X^2}\right) + b\left(x - m_X\right)\right) f(x,y)dx\,dy}_{=a^2\sigma_{X^2}^2 + b^2\sigma_X^2} +$$

$$+ \underbrace{\int_{-\infty}^{\infty} \left(a\left(x^2 - m_{X^2}\right) + b\left(x - m_X\right)\right)^2 f(x)dx}_{=a^2\sigma_{X^2}^2 + b^2\sigma_X^2} \underbrace{\int_{-\infty}^{\infty} f(y|x)dy}_{=1} =$$

$$= \sigma_Y^2 - a^2\sigma_{X^2}^2 - b^2\sigma_X^2 = \sigma_Y^2\left(1 - r_{X^2Y}^2 - r_{XY}^2\right) < \sigma_Y^2\left(1 - r_{XY}^2\right). \qquad (14)$$

При этом неравенство (11) приобретает вид

$$\eta^2 = D\left(m_Y(x)\right)\Big/\sigma_Y^2 \geq r_{X^2Y}^2 + r_{XY}^2, \qquad (15)$$

причем равенство возможно только в ситуации точного равенства (13), когда корреляция квадратична по сути.

В общем случае учет нелинейности корреляционной зависимости лишь усиливает неравенство, подтверждающее адекватность регрессионного приближения (см. (12)

$$\left(\eta^2 - r_{X^2Y}^2 - r_{XY}^2\right)\Big/\eta^2 < \left(\eta^2 - r_{XY}^2\right)\Big/\eta^2 \ll 1. \qquad (16)$$

Если и этого недостаточно (например, график условного матожидания содержит точку перегиба) целесообразно заменить в квадратный полином на кубический, и т. д. Сходимость такой процедуры (если $m_Y(x)$ не полином по сути) вытекает из того, что правая часть (16) неотрицательна по смыслу.

ЛИТЕРАТУРА

1. Гмурман, В. Е. Теория вероятностей и математическая статистика. – М. : Высшая школа, 1972. – 368 с.
2. Коваленко, И. Н., Филиппова, А. А. Теория вероятностей и математическая статистика. – М. : Высшая школа, 1973. – 368 с.

3. Ekhilevskiy, S. G., Golubeva, O. V., Zabelendik, O. N., Struk, T. S. Contribution of excesses / of the gamma distribution to the asymptotics of factorials with large arguments // System analysis and information Technology. – 2019. – № 1(15)–2(16). – С. 119–125.
4. Ехилевский С. Г., Голубева О. В., Потапенко Е. П. Теоретико-вероятностный подход к моделированию респиратора на химически связанном кислороде // Безопасность труда в промышленности.– 2020. – № 10. – С. 7–15.
5. Ехилевский С. Г., Голубева О. В., Потапенко Е. П., Рудькова Т. С. Независимые повторные испытания как асимптотически гауссовский случайный процесс // Вестник Полоцкого государственного университета. Серия С. Фундаментальные науки. – 2016. – № 2, с. 111–116.