Лекция 10

Статистический анализ зависимостей между гидрологическими переменными

Метод наименьших квадратов
Уравнение линейной регрессии для двух переменных
Линерализация нелинейных зависимостей
Оценка точности уравнения линейной регрессии для двух переменных

(Ахметов С.К.)

Статистический анализ зависимостей между гидрологическими переменными

Задача: Найти вид зависимости $y = f(x_1, x_2, \ldots x_k)$
где y - зависимая переменная (или предиктант)
$x_1, x_2, \ldots x_k$ – независимые переменные (предикторы)

Допустим для простоты, что y зависит только от одного предиктора, т.е. $y = f(x)$ и что зависимость $y = f(x)$ является линейной

Искомым уравнением регрессии в этом случае будет выражение

$$y_i = ax_i + b$$

Метод наименьших квадратов

- Нужно определить такие значения параметров ***a*** и ***b***, при которых сумма квадратов отклонений наблюденных значений y_i от рассчитанных по вышеприведенной формуле будет иметь минимальное значение.
- Сумма квадратов отклонений равна

$$\sum_{i=1}^{n}(y_i-\bar{y}_i)^2=\sum_{i=1}^{n}\left[y_i-(ax_i+b)\right]^2$$

Чтобы сумма стала минимальной частные производные по параметрам ***a*** и ***b*** должны равняться нулю.

$$\frac{\partial}{\partial a}\left[\sum_{i=1}^{n}(y_i-ax_i-b)^2\right]=0;\qquad \frac{\partial}{\partial b}\left[\sum_{i=1}^{n}(y_i-ax_i-b)^2\right]=0.$$

Метод наименьших квадратов

Решая эти уравнения относительно ***a*** и ***b***, получим

$$a = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

$\bar{x}$ и $\bar{y}$ - средние значения ***X*** и ***Y***

$$a_{y/x} = r\frac{\sigma_y}{\sigma_x}$$

a - коэффициент регрессии. Он равен

σ_y и σ_x - среднеквадратические отклонения выборок из ***Y*** и ***X***

r – выборочный коэффициент парной корреляции, определяемый по формуле

$$r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Существует связь между коэффициентом корреляции и параметрами регрессии

$$r_{yx} = \sqrt{a_{y/x}a_{x/y}}$$

Метод наименьших квадратов

r - эмпирическая мера линейной зависимости между ***Y*** и ***X,*** изменяется от -1 до +1. При знаке «+» - зависимость прямая, а при знаке «-« - обратная

Коэффициент корреляции можно рассчитать по формуле

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}.$$

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

с учетом того, что

$$b = \bar{y} - a\bar{x}$$

$$a_{y/x} = r\frac{\sigma_y}{\sigma_x}$$

уравнение регрессии можно представить в виде

$$\left(y_i - \bar{y}\right) = r\frac{\sigma_y}{\sigma_x}\left(x_i - \bar{x}\right)$$

Линеаризация нелинейных зависимостей

Зависимость $y = f(x)$ может иметь и нелинейных вид

В этом случае, можно попытаться использовать для аппроксимации зависимости $y = f(x)$ уравнение экспоненты

$$y = c\, e^{ax}$$

где a и c - эмпирические параметры

Метод наименьших квадратов позволяет определить параметры и в случае нелинейной модели. Можно существенно упростить расчеты, проведя линеаризацию исходного выражения.

Прологарифмировав обе части получим

$$ln(y) = ln(c) + ax$$

Обозначим y и x в виде $y' = ln(y);\ x' = x$. С учетом этого перепишем выражение выше $y' = ax' + b$, где $b = ln\,(c)$

Теперь уравнение стало линейным и для оценки a и b можно использовать подход, который использовался в первом случае.

После того как параметры найдены, проводят обратное преобразование. В данном случае: $c = e^b$

Преобразования, применяемые при линеаризации зависимостей

Исходная зависимость	Преобразование		Уравнение в линейной форме
	Абсцисса	Ордината	
$y = c\,e^{ax}$	x	$\ln(y)$	$\ln(y) = ax + \ln(c)$
$y = c\,x^{a}$	$\ln(x)$	$\ln(y)$	$\ln(y) = a\ln(x) + \ln(c)$
$y = (a/x) + b$	$1/x$	y	$y = a(1/x) + b$
$y = x/(ax + b)$	x	x/y	$(x/y) = ax + b$
$y = c/(kx + m)$	x	$1/y$	$(1/y) = (k/c)\,x + (m/c)$

Оценка точности уравнения линейной регрессии для двух переменных

- Обычно в гидрологии регрессионная зависимость может использоваться для практических расчетов, если $|r| \geq 0.7$

Другие статистические характеристики, позволяющие судить о точности полученного уравнения

$\sigma_{y(x)}$ – ***стандартная ошибка уравнения линейной регрессии.*** Эта величина характеризует среднеквадратическое отклонение точек от принятой линии регрессии.

$$\sigma_{y(x)} = \sqrt{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 / (n-2)}$$

y_i – наблюденная величина

$\tilde{y}_i$ - величина, рассчитанная по уравнению регрессии

n-2 - число степеней свободы.

Число степеней свободы равно числу наблюдений минус число параметров, определяемых по эмпирическим данным. В данном случае таких параметров два: коэффициент регрессии ***a*** и свободный член – ***b***.

Оценка точности уравнения линейной регрессии для двух переменных

$\sigma_{y(x)}$ через коэффициент корреляции можно записать

$$\sigma_{y(x)} = \sigma_y^* \sqrt{\frac{(1-r^2)(n-1)}{(n-2)}}$$

где σ_y^* - несмещенная оценка СКО для ряда ***Y***

Иногда при практических расчетах пренебрегают величиной $\sqrt{(n-1)/(n-2)}$ и используют более простую формулу

$$\sigma_{y(x)} = \sigma_y^* \sqrt{1-r^2}$$

Оценка точности уравнения линейной регрессии для двух переменных

σ_r - *стандартная ошибка коэффициента парной корреляции*

$$\sigma_r = \left(1 - r^2\right) / \sqrt{n-1}$$

σ_a – *стандартная ошибка коэффициента регрессии*

$$\sigma_a = \sigma_{y(x)} \Big/ \sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

Выражение для σ_a можно представить также в виде

$$\sigma_a = \frac{\sigma_y^*}{\sigma_x^*} \sqrt{\frac{1 - r^2}{n-2}}$$

где σ_x^* и σ_y^* - оценки СКО соответственно для X и Y

Оценка точности уравнения линейной регрессии для двух переменных

σ_b – *стандартная ошибка свободного члена*

$$\sigma_b = \sigma_{y(x)} \sqrt{\sum_{i=1}^{n} (x_i)^2 \Big/ n \sum_{i=i}^{n} (x_i - \bar{x})^2}$$

или это выражение еще можно записать как

$$\sigma_b = \sigma_y^* \sqrt{\frac{1-r^2}{n-2}} \sqrt{1 + \left(\frac{\bar{x}}{\sigma_x^*}\right)^2}$$

Для практических расчетов можно рекомендовать следующие соотношения, при которых можно использовать уравнения регрессии

$$n \geq 10; \quad |r| \geq 0{,}7; \quad |r| / \sigma_r \geq 2; \quad |a| / \sigma_a \geq 2$$

Желательное, но необязательное условие

$$\frac{|b|}{\sigma_b} \geq 2$$

СПАСИБО ЗА ВНИМАНИЕ!