Лекция 11

Статистический анализ зависимостей между гидрологическими переменными

Интервальная оценка и оценка значимости параметров линейной регрессии для двух переменных. Интервальная оценка коэффициента парной корреляции. Коэффициент ранговой корреляции Спирмэна. Интервальная оценка коэффициента регрессии. Интервальная оценка свободного члена

(Ахметов С.К.)

Интервальная оценка и оценка значимости параметров линейной регрессии для двух переменных

В случае, если r не очень велико и длина выборок не превышает 40 лет, то распределение коэффициентов корреляции хорошо аппроксимируется нормальным законом со среднеквадратическим отклонением $\sigma^*_{r'}$

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

доверительный интервал для истинного коэффициента корреляции можно представить в виде

$$r^* - t'_{1-\alpha}\sigma_r^* \le r < r^* + t'_{1-\alpha}\sigma_r^*$$

где r^* - выборочный коэффициент парной корреляции

$t'_{1-\alpha}$ – квантиль стандартного нормального распределения, соответствующий двустороннему уровню значимости 2α

Z – преобразование Фишера

В случае, если значения ***r>0.4*** и ***n <40***, для построения доверительного можно использовать ***Z*** – преобразование Фишера, которое связано с ***r*** выражением

$$Z = 0.5 \ln[(1+ r)/(1- r)]$$

В отличие от ***r*** статистика ***Z*** имеет нормальное распределение даже при ***n*** небольшом. СКО для ***Z*** определяется по формуле

$$\sigma_r = 1/\sqrt{n-3}$$

Последовательность построения интервальной оценки *r* при использовании преобразования Фишера

1. Рассчитывается Z по формуле $Z = 0.5 \ln[(1+ r)/(1- r)]$

2. Рассчитывается **СКО** Z по формуле

$$\sigma_r = 1/\sqrt{n-3}$$

3. Строиться доверительный интервал для Z

$$Z^* - t'_{1-\alpha}\sigma_z^* \le r < Z^* + t'_{1-\alpha}\sigma_z^*$$

4. Строиться доверительный интервал для коэффициента корреляции путем обратного перехода от Z к r, т.е.

$$(e^{2z'} - 1)/(e^{2z'} + 1) \le r < (e^{2z''} - 1)/(e^{2z''} + 1)$$

Здесь $z' = Z^* - t'_{1-\alpha}\sigma_z^*$ и $z'' = Z^* + t'_{1-\alpha}\sigma_z^*$

Z – преобразование Фишера точнее и может быть рекомендовано при любых значениях $r>0.4$ и $n<40$

Проверки значимости линейной зависимости между *X* и *Y*

Коэффициент корреляции можно использовать для проверки значимости линейной зависимости между ***X*** и ***Y***.

В этом случае выдвигается нулевая гипотеза, что ***r=0,*** т.е. что связь полностью отсутствует.

Гипотеза опровергается, если

$$\frac{|r^*|}{\sigma_r^*} > t'_{1-\alpha}$$

и связь считается статистически значимой.

Если это условие не выполняется, то связь статистически незначима.

Следует иметь в виду, что здесь имеется в виду двусторонний уровень значимости, т.е. вы задаете чему равно 2α, а потом находите α. Допустим, что $2\alpha = 5\%$, тогда $t'_{1-\alpha} = t_{97,5}$.

Коэффициент ранговой корреляции Спирмэна

$$r_S = 1 - \frac{6\sum_{i=1}^{i=n} \Delta_i^2}{n(n^2-1)}$$

Если распределение случайных рядов y_1, $y_2 \ldots y_n$ и $x_1, x_2 \ldots x_n$ существенно отличается от нормального распределения, то для оценки степени их взаимосвязанности можно использовать коэффициент ранговой корреляции Спирмэна r_s:

где n – длина выборок;

Δ_i – разность рангов для пары значений y_i и x_i

Для коэффициента ранговой корреляции выполняется условие:

$$-1 \le r_s \le +1$$

Выдвигается нулевая гипотеза о том, что $r_s = 0$

Гипотеза опровергается, если $\left|r_S^*\right| > (r_S)_\alpha$

$(r_s)_\alpha$ – критическое значение коэффициента ранговой корреляции при одностороннем уровне значимости α

Для $n \le 30$ значение $(r_s)_\alpha$ представлены в таблице

Критические значения коэффициента ранговой корреляции Спирмэна,

n	Уровень значимости, α %				
	0,5	1	2,5	5	10
5		0,9000	0,9000	0,8000	0,7000
6	0,9429	0,8857	0,8286	0,7714	0,6000
7	0,8929	0,8571	0,7450	0,6786	0,5357
8	0,8571	0,8095	0,6905	0,5952	0,4762
9	0,8167	0,7667	0,6833	0,5833	0,4667
10	0,7818	0,7333	0,6364	0,5515	0,4424
11	0,7545	0,7000	0,6091	0,5273	0,4182
12	0,7273	0,6713	0,5804	0,4965	0,3986
13	0,6978	0,6429	0,5549	0,4780	0,3791
14	0,6747	0,6220	0,5341	0,4593	0,3626
15	0,6536	0,6000	0,5179	0,4429	0,3500
16	0,6324	0,5824	0,5000	0,4265	0,3382
17	0,6152	0,5637	0,4853	0,4118	0,3260
18	0,5975	0,5480	0,4716	0,3994	0,3148
19	0,5825	0,5333	0,4579	0,3895	0,3070
20	0,5684	0,5203	0,4451	0,3789	0,2977
21	0,5545	0,5078	0,4351	0,3688	0,2909
22	0,5426	0,4963	0,4241	0,3597	0,2829
23	0,5306	0,4852	0,4150	0,3518	0,2767
24	0,5200	0,4748	0,4061	0,3435	0,2704
25	0,5100	0,4654	0,3977	0,3362	0,2646
26	0,5002	0,4564	0,3894	0,3299	0,2588
27	0,4915	0,4481	0,3822	0,3236	0,2540
28	0,4828	0,4401	0,3749	0,3175	0,2490
29	0,4744	0,4320	0,3685	0,3113	0,2443
30	0,4665	0,4251	0,3620	0,3059	0,2400

Коэффициент ранговой корреляции Спирмэна

При $n \geq 30$ величина $r_s\sqrt{(n-1)}$ достаточно хорошо описывается нормальным распределением. В этом случае нулевая гипотеза $(r_s=0)$ отвергается , если выполняется неравенство

$$\left|r_S^*\right| > \frac{t'_{1-\alpha}}{\sqrt{n-1}}$$

где $t'_{1-\alpha}$ – квантиль стандартного нормального распределения при одностороннем уровне значимости α.

Последовательность расчетов по методу коэффициента ранговой корреляции Спирмэна

1. Ряды y_i и x_i ранжируются в возрастающем порядке
2. Каждому значению y_i и x_i в ранжированном ряду присваивается порядковый номер (ранг). Самое маленькое значение случайной величины получает первый ранг и т.д.
3. Каждому значению случайной величины ставится свой ранг
4. Рассчитывается разность рангов y_i и x_i
5. Рассчитывается квадрат разности рангов Δ^2
6. По формуле ниже рассчитывается коэффициент ранговой корреляции

$$r_S = 1 - \frac{6\sum_{i=1}^{i=n} \Delta_i^2}{n(n^2-1)}$$

7. По таблице опред-ся критический коэффициент ранговой корреляции
8. Выдвигается нулевая гипотеза о том, что $r_s = 0$

Гипотеза опровергается, если $\left|r_S^*\right| > (r_S)_\alpha$

Расчет коэффициента ранговой корреляции

№ п/п	Исходные ряды		Ранжированные ряды		Ранг		Δ	Δ^2
	y_i	x_i	y_i	x_i	y_i	x_i		
1	15,2	79,0	8,32	40,4	16	18	–2	4
2	11,6	59,5	9,29	46,6	9	8	1	1
3	17,3	84,3	10,4	47,6	22	22	0	0
4	20,7	95,5	10,5	48,2	29	26	3	9
5	10,4	46,6	10,9	53,2	3	2	1	1
6	11,3	48,2	11,3	54,0	6	4	2	4
7	16,3	69,0	11,4	54,6	19	11	8	64
8	22,3	105	11,5	59,5	30	30	0	0
9	16,1	75,0	11,6	61,6	18	15	3	9
10	14,9	78,8	12,8	66,1	15	17	–2	4
11	15,6	72,3	13,0	69,0	17	13	4	16
12	10,5	53,2	13,8	69,9	4	5	–1	1
13	13,0	54,0	14,3	72,3	11	6	5	25
14	8,32	40,4	14,3	72,5	1	1	0	0
15	9,29	47,6	14,9	75,0	2	3	–1	1
16	12,8	66,1	15,2	77,6	10	10	0	0
17	11,4	54,6	15,6	78,8	7	7	0	0
18	14,3	69,9	16,1	79,0	13	12	1	1

Интервальная оценка коэффициента регрессии

Если разброс наблюдений относительно линейной регрессии нормален, то доверительный интервал для коэффициента регрессии имеет вид

$$a^* - t_{1-\alpha}\sigma_a < a \le a^* + t_{1-\alpha}\sigma_a$$

где a^* - эмпирической значение коэффициента регрессии

σ_a – стандартная ошибка коэффициента регрессии

$t'_{1-\alpha}$ – квантиль распределения Стьюдента, соответ-щий двухстороннему уровню значимости 2α при числе степеней свободы $v = n - 2$

При проверке значимости коэффициента регрессии выдвигается нулевая гипотеза о том, что $a=0$. Гипотеза опровергается, если

$$\left| t_a^* \right| > t_{1-\alpha}$$

t^*_a – эмпирическое значение статистики Стьюдента, определяемое по формуле

$$t_a^* = a^* / \sigma_a$$

Если равенство выполняется, то коэффициент регрессии считается статистически значимым, в противном случае коэффициент a^* является статистически незначимым и линейная связь между X и Y отсутствует.

Интервальная оценка свободного члена

Доверительный интервал для свободного члена имеет вид

$$b^* - t_{1-\alpha}\sigma_b < b \le b^* + t_{1-\alpha}\sigma_b$$

где b^* - эмпирической значение коэффициента регрессии

σ_b – стандартная ошибка коэффициента регрессии

$t'_{1-\alpha}$ – квантиль распределения Стьюдента, соответ-щий двухстороннему уровню значимости 2α при числе степеней свободы $v = n - 2$

При проверке значимости коэффициента регрессии выдвигается нулевая гипотеза о том, что $b=0$. Гипотеза опровергается, если

$$\left| t_b^* \right| > t'_{1-\alpha}$$

t^*_b – эмпирическое значение статистики Стьюдента, определяемое по формуле

$$t_b^* = b^* / \sigma_b$$

Если равенство выполняется, то коэффициент регрессии считается статистически значимым, в противном случае коэффициент b^* является статистически незначимым и для аппроксимации зависимости между X и Y вместо выражения

$\tilde{y}_i = ax_i + b$ следует использовать выражение $\tilde{y}_i = ax_i$

F – критерий значимости регрессии

Часто для проверки значимости линейной регрессии используется критерий

$$F^*_{y(x)} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \Big/ (n-2)} = \frac{(n-1)S_y^2 r^2}{\sigma^2_{y(x)}}$$

Доказано, что это отношение имеет распределение Фишера со степенями свободы $v_1 = 1$ и $v_2 = n-2$. Связь считается значимой, если

$$F^*_{y(x)} > F_{1-\alpha}$$

где $F_{1-\alpha}$ – теоретическое значение статистики Фишера при уровне значимости α

Построение доверительного интервала для уравнения регрессии

Доверительные пределы для уравнения регрессии определяются по формуле

$$\tilde{y}_{k,0} = \tilde{y}_k \pm t'_{1-\alpha}\,\sigma_{регр}$$

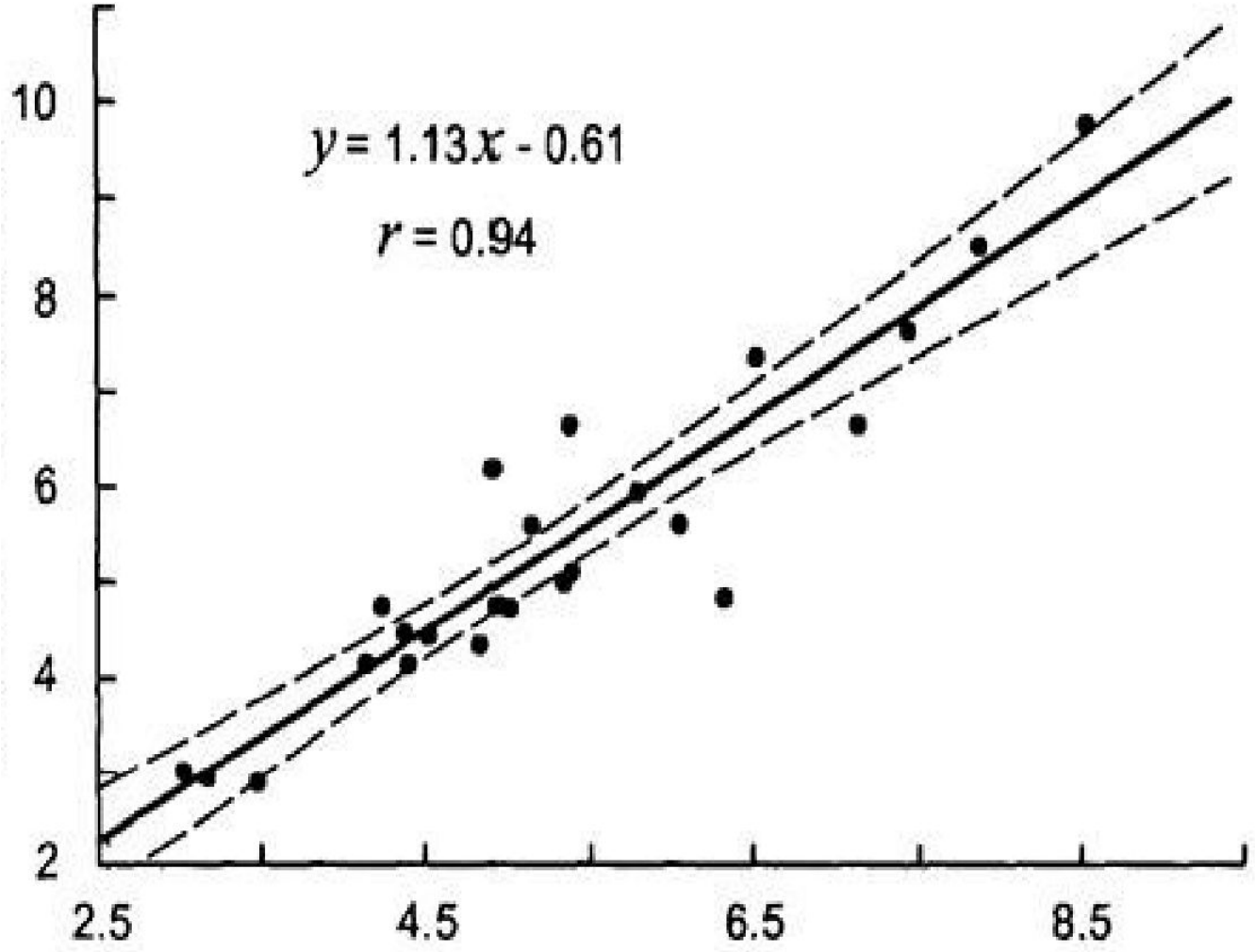
$\tilde{y}_{k,0}$ - истинное значение случайной величины

$\tilde{y}_k = ax_k + b$ - это расчетное значение функции

t' $_{1-\alpha}$ – квантиль распределения Стьюдента, соответствующее двухстороннему уровню значимости ***2α*** при числе степеней свободы ***ν = n-2***

$$\sigma_{регр} = \sigma_{y(x)}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$\sigma_{y(x)}$ - стандартная ошибка уравнения линейной регрессии


y = 1.13x - 0.61
r = 0.94
10
8
6
4
2
2.5
4.5
6.5
8.5

СПАСИБО ЗА ВНИМАНИЕ!