Регрессионный анализ материалов наблюдений

Цель задания: освоение технологии проведения регрессионного анализа.

Исходные данные: парные значения x_i и y_i $i = 1,2,...,n$ случайных величин X и Y, представляющие состояние системы (X, Y) в разные моменты ее наблюдения.

План

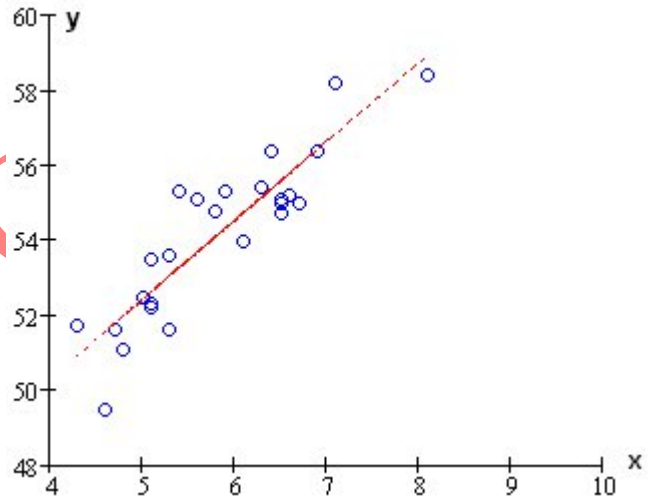
1. Построить поле корреляции - точечную диаграмму (график).
2. На основании поля корреляции выдвинуть гипотезу о линейной корреляционной зависимости величин X Equation.DSMT4 и Y с функцией регрессии $M\left(Y/_{X=x}\right) = Ax + B$.
3. Вычислить оценки $\overline{X}$ и $\overline{Y}$ математических ожиданий случайных величин X Equation.DSMT4 и Y.
4. Оценить средние квадратические отклонения σ_X и σ_Y случайных величин X Equation.DSMT4 и Y.
5. Вычислить выборочный коэффициент корреляции r_{XY} - оценку действительного коэффициента корреляции r_{XY}.
6. Проверить гипотезу о не значимости выборочного коэффициента корреляции.
7. Получить оценку функции регрессии - уравнение регрессии Y на X Equation.DSMT4 . Записать уравнение регрессии *в числовом виде*. Нанести прямую регрессии на график.
8. Оценить точность регрессии (*точность прогнозов*).
9. Выполнить точечную и интервальную оценку точности коэффициентов a и b уравнения регрессии.
10. Оценить значимость коэффициента a Equation.DSMT4 уравнения регрессии.
11. Оценить точность и надежность выборочного коэффициента корреляции согласно критерию Фишера.
12. Сделать вывод *о наличии, тесноте и форме* корреляционной зависимости между случайными величинами X и Y.

Указания: 1. Проверку всех гипотез выполнить на 5%-м уровне значимости (т. е. q=0.05 *и* $\beta = 0.95$*).*

2. Необходимые вычисления выполнить в рабочей таблиц (в Excel).

Решение

1. Точечная диаграмма (поле корреляции)



2. На основании графика выдвигаем гипотезу о линейной связи величин *X* и *Y:*

Образец оформления

$$H_0 = \{ Y = A \times X + B \}$$

Рабочая таблица

i	x_i	y_i	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})^2$	$(x_i - \bar{X}) \times (y_i - \bar{Y})$	$\hat{y}_i$	$(\hat{y}_i - y_i)^2$
1	6,3	55,4	0,472	1,248	0,223	1,558	0,589	55,150	0,063
2	5,3	51,6	-0,528	-2,552	0,279	6,513	1,347	53,036	2,061
3	6,5	54,7	0,672	0,548	0,452	0,300	0,368	55,573	0,762
4	5,8	54,8	-0,028	0,648	0,001	0,420	-0,018	54,093	0,500
5	5,1	52,2	-0,728	-1,952	0,530	3,810	1,421	52,613	0,170
6	8,1	58,4	2,272	4,248	5,162	18,046	9,651	58,956	0,309
7	4,8	51,1	-1,028	-3,052	1,057	9,315	3,137	51,979	0,772
8	4,3	51,7	-1,528	-2,452	2,335	6,012	3,747	50,921	0,606
9	6,4	56,4	0,572	2,248	0,327	5,054	1,286	55,361	1,079
10	4,6	49,5	-1,228	-4,652	1,508	21,641	5,713	51,556	4,226
11	5,6	55,1	-0,228	0,948	0,052	0,899	-0,216	53,670	2,045
12	6,9	56,4	1,072	2,248	1,149	5,054	2,410	56,418	0,000
13	5,4	55,3	-0,428	1,148	0,183	1,318	-0,491	53,247	4,214
14	5,1	52,3	-0,728	-1,852	0,530	3,430	1,348	52,613	0,098
15	5,1	53,5	-0,728	-0,652	0,530	0,425	0,475	52,613	0,787
16	7,1	58,2	1,272	4,048	1,618	16,386	5,149	56,841	1,846
17	5	52,5	-0,828	-1,652	0,686	2,729	1,368	52,401	0,010
18	4,7	51,5	-1,128	-2,652	1,272	7,033	2,991	51,767	0,071
19	6,6	55,2	0,772	1,048	0,596	1,098	0,809	55,784	0,341
20	5,3	53,6	-0,528	-0,552	0,279	0,305	0,291	53,036	0,318
21	6,5	55,1	0,672	0,948	0,452	0,899	0,637	55,573	0,224
22	6,7	55	0,872	0,848	0,760	0,719	0,739	55,996	0,991
23	6,5	55	0,672	0,848	0,452	0,719	0,570	55,573	0,328
24	6,1	54	0,272	-0,152	0,074	0,023	-0,041	54,727	0,529
25	5,9	55,3	0,072	1,148	0,005	1,318	0,083	54,304	0,992
S	145,7	1353,8	0,000	0,000	20,510	115,022	43,364	1353,800	23,342

$\bar{X} =$	$\bar{Y} =$	$\tilde{s}_X =$	$\tilde{s}_Y =$	$r_{XY} =$	$\tilde{s}_{ост.} =$
5,828	54,152	0,906	2,14	0,86	1,01

3. Оценка *M(X)*: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = 5,828$ – среднее арифметическое;

   Оценка *M(Y)*: $\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} = 54,152$ – среднее арифметическое;

4. Оценка ср. кв. отклонения СВ. *X*: $\sigma_X = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n-1}} = 0.906;$

   Оценка ср. кв. отклонения СВ. *Y*: $\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{Y})^2}{n-1}} = 2,14;$

5. Выборочный коэффициент корреляции: $r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{X}) \times (y_i - \bar{Y})}{n \times \sigma_X \times \sigma_Y} = 0,86.$

Образец оформления

6. Нулевая гипотеза о не значимости коэффициента корреляции: $H_0 = \{ r_{XY} = 0 \}$. Эмпирическое значения критерия проверки $t_э = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \approx 7{,}98$. Критическое значение $t_T \approx 2{,}07$ - из таблиц распределения Стьюдента для доверительной вероятности $\beta = 0.95$ и числа степеней свободы $n = n - 2 = 23$. Так как $t_э > t_T$, то нулевая гипотеза отклоняется и коэффициент корреляции значим.

7. Коэффициенты уравнения регрессии $\hat{y} = a \cdot x + b$ Equation.DSMT4 (*формулы прогнозов*):

$a = r \cdot \frac{\sigma_Y}{\sigma_X} = 2{,}03; \qquad b = \bar{Y} - a \cdot \bar{X} = 42{,}3.$

Уравнение регрессии в числовом виде: $\hat{y} = 2{,}03 \cdot x + 42{,}3$. Наносим прямую линию на график поля корреляции.

8. Оценка точности регрессии (*точность прогнозов по уравнению регрессии*):

$s_{ост.} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i)^2}{n-2}} = 1{,}01$ – остаточное ср.кв. отклонение.

9. Оценка точности коэффициентов уравнения регрессии.

а) точечная:

$\sigma_a = \frac{s_{ост.}}{\sigma_X\sqrt{(n-2)}} = 0{,}23; \quad \sigma_b = \frac{s_{ост.}}{\sqrt{(n-2)}} = 0{,}21$ - ср. кв. отклонения коэффициентов a Equation.DSMT4 и b.

б) интервальная:

$P\left(a - t_\beta \cdot \sigma_a \le A \le a + t_\beta \cdot \sigma_a\right) = \beta$ - доверительный интервал для коэффициента A функции регрессии;

$P\left(b - t_\beta \cdot \sigma_b \le B \le b + t_\beta \cdot \sigma_b\right) = \beta$ - доверительный интервал для коэффициента B функции регрессии.

$t_\beta \approx 2.0$ - аргумент функции Лапласа, соответствующий принятой доверительной вероятности $\beta = 0.95$.

Тогда: $P\left(2{,}03 - 2{,}0 \cdot 0{,}23 \le A \le 2{,}03 + 2{,}0 \cdot 0{,}23\right) = \beta$, $P\left(42{,}3 - 2{,}0 \cdot 0{,}21 \le B \le 42{,}3 + 2{,}0 \cdot 0{,}21\right) = \beta$ и окончательно: $P\left(1{,}77 \le A \le 2{,}49\right) = 0{,}95$, $\left(41{,}88 \le B \le 42{,}72\right) = 0{,}95$.

10. Для проверки значимости коэффициента a Equation.DSMT4 уравнения регрессии проверим нулевую гипотезу $H_0 = \{ A = 0 \}$ о не значимости коэффициента A функции регрессии.

Эмпирическое значение критерия проверки гипотезы $t_э = \left|\frac{a}{\sigma_a}\right| = 8{,}9$. Критическое значение $t_T \approx 2{,}07$ - из таблиц распределения Стьюдента для доверительной вероятности $\beta = 0{,}95$ и числа степеней свободы $n = n - 2 = 23$. Так как $t_э > t_T$, то нулевая гипотеза отклоняется и коэффициент a уравнения регрессии значим.

11. Для оценки точности и надежности выборочного коэффициента корреляции построим доверительный интервал для теоретического коэффициента корреляции r_{XY} согласно критерию Фишера:

$$P\left(thZ_1 \le r_{XY} \le thZ_2\right) = \beta,$$

где *th* - символ гиперболического тангенса;

$r_{XY} = 0{,}86, \quad Z = \frac{1}{2}\ln\left(\frac{1 + r_{XY}}{1 - r_{XY}}\right) = 1{,}28, \quad s_Z = \frac{1}{\sqrt{n-3}} = 0{,}213;, \quad t_\beta = 2{,}07$ - аргумент функции Лапласа $\Phi(t_\beta) = \beta$;

$t_\beta \cdot s_Z = 0.43; \quad Z_1 = Z - t_\beta \cdot s_Z = 0{,}85; \quad Z_2 = Z + t_\beta \cdot s_Z = 1{,}71; \; thZ_1 = 0{,}691; \quad thZ_2 = 0{,}937.$

Окончательно: $P(0{,}691 \le r_{XY} \le 0{,}937) = 0{,}95$ - доверительный интервал для r_{XY}.

$\left(thZ_2 - thZ_1\right) = 0{,}246, \qquad r_{min} = \left(\sqrt{n+36} - \sqrt{n}\right)/6 = 0{,}67.$

Так как длина доверительного интервала $\left(thZ_2 - thZ_1\right) < |r_{XY}|$, равно как и $r_{\min} < thZ_1$, то значение выборочного коэффициента корреляции следует считать надежным и линейную корреляционную зависимость – установленной.

Вывод: между случайными величинами X Equation.DSMT4 и Y существует линейная корреляционная зависимость с коэффициентом корреляции $r_{XY} = 0{,}86$ в форме уравнения регрессии $\hat{y} = (2{,}03 \cdot x + 42{,}3) \pm 1{,}01$. Точность прогнозов по уравнению регрессии $s_{ост.} = 1{,}01$.

Образец оформления