ЛР № 4. Анализ данных в линейной регрессионной модели.

X	Y
16,41	15,47
12,61	11,76
10,29	10,21
17,19	20,35
5,6	4,71
13,85	18,37
13,62	16,23
18,27	20,94
6,09	6,71
0,64	6,19
9,93	12,11
15,5	15,9
18,11	14,6
8,38	10,55
0,93	–0,91
19,47	18,42
4,78	6,86
16,59	14,51
20,17	19,2
12,2	13,87
17,59	17,62
8,2	8,42
17,14	15,25
17,09	14,34
6,2	5,5
8,56	5,13
0,44	–1,88
8,97	10,71
17,79	18,96
1,39	–5,16
11,75	11,91
–0,55	–0,68
5,98	1,51
10,44	12,04
12,15	13,07
11,98	12,84
4,09	6,2
9,74	9,54
0,95	0,87
12,21	11,72
6,31	6,83
8,34	5,48
13,24	15,45
5,78	3,64
5,03	6,26
9,85	5,69
4,53	9,83
12,46	10,01
11,73	10,06
7,79	9,86

1. Для негруппированных данных проверить гипотезу $H_0: \rho_{X,Y} = 0$ об отсутствии линейной статистической связи между компонентами X и Y при альтернативной гипотезе $H_1: \rho_{X,Y} \neq 0$ (уровень значимости $\alpha = 0.05$)

```
clc; clear;
X = [16.41 12.61 10.29 17.19 5.60 13.85 13.62 18.27 6.09 0.64 9.93 15.50 18.11 8.38
0.93 19.47 4.78 16.59 20.17 12.20 17.59 8.20 17.14 17.09 6.20 8.56 0.44 8.97 17.79
1.39 11.75 -0.55 5.98 10.44 12.15 11.98 4.09 9.74 0.95 12.21 6.31 8.34 13.24 5.78
5.03 9.85 4.53 12.46 11.73 7.79];
Y = [15.47 11.76 10.21 20.35 4.71 18.37 16.23 20.94 6.71 6.19 12.11 15.90 14.60 10.55
-0.91 18.42 6.86 14.51 19.20 13.87 17.62 8.42 15.25 14.34 5.50 5.13 -1.88 10.71 18.96
-5.16 11.91 -0.68 1.51 12.04 13.07 12.84 6.20 9.54 0.87 11.72 6.83 5.48 15.45 3.64
6.26 5.69 9.83 10.01 10.06 9.86];
alpha = 0.05;
%%
n = max(size(X)); % n = 50
corr_XY_matrix = corrcoef(X, Y); %корр. матрица
q_XY = corr_XY_matrix(1,2); % q_XY = 0.9154

%Найдём квантиль распределения Стьюдента t_(1 - alpha / 2)

t_0975 = tinv(1-alpha/2, n-2) % t_0975 от (n-2) = 2.0106
%Тогда выборочное значение статистики Z равно:
Z_s = q_XY * sqrt(n-2) / sqrt(1 - q_XY^2) % Z_s = 15.7575
if(abs(Z_s) > t_0975)
    disp('Гипотеза H0 отклоняется в пользу H1')
else
    disp('Гипотеза H0 принимается (H1 отклоняется)')
end % |15.7575| > 2.0106 => H0 отклоняется в пользу H1
% => Линейная статистическая связь между компонентами X и Y присутствует!
```

2. для негруппированных данных получить интервальную оценку для истинного значения коэффициента корреляции $\rho_{X,Y}$ при уровне значимости $\alpha = 0.05$;

$$\text{Так как } |Z_в| > t_{0.975}(48),$$
$$\text{то гипотеза } H_0 \text{ отклоняется в пользу } H_1. \text{ Корреляция значима.}$$

Менее чувствительной к объёму n выборки из генеральной совокупности, имеющей двумерное нормальное распределение, является статистика U для проверки более общей гипотезы $H_0: \rho_{X,Y} = \rho_0$ против любой из трёх альтернатив

$$H_1{}^{(1)}: \rho_{X,Y} < \rho_0 \; ; \; H_1{}^{(2)}: \rho_{X,Y} \neq \rho_0; \; H_1{}^{(3)}: \rho_{X,Y} > \rho_0.$$

$$U = \frac{\left(arth\left(\tilde{\rho_{X,Y}}\right) - arth(\rho_0)\right)}{\frac{1}{\sqrt{(n-3)}}},$$

где $arth(x) = \frac{1}{2}\ln\frac{1+x}{1-x}$ $(|x| < 1)$, а ρ_0 – истинное неизвестное значение

При $n \geq 30$ $U \sim N(0;1)$

Критерий проверки $H_0: \rho_{X,Y} = \rho_0, H_1: \rho_{X,Y} \neq \rho_0$ *заключается в следующем:*
-вычисляется выборочное значение $u_в$ статистики U

-если $|u_в| < u_{1-\frac{\alpha}{2}}$, *то нет оснований отвергать гипотезу* H_0; *иначе* H_0 *отклоняется с ошибкой первого рода* α *в пользу* $H_1{}^{(2)}$

```
%%
clc;
% для негруппированных данных получить интервальную оценку для истинного значения
% коэффициента корреляции ρ_(X,Y) при уровне значимости α=0.05
% Менее чувствительной к объёму n выборки из генеральной совокупности, имеющей
двумерное
% нормальное распределение, является статистика U для проверки более общей
% гипотезы H0:  ρ_(X,Y) =  ρ_0.

%Воспользуемся статистикой Фишера U
%u_1-alpha/2 = u_0.975 = 1,96
u_0975 = 1.96 %finv(1-alpha/2, n, n)

roeInt = [0 0];
roeInt(1) = atan(0.5 * log((1 + q_XY) / (1 - q_XY)) - (u_0975) / (sqrt(n-3)) - q_XY /
(2* (n-1)) );
roeInt(2) = atan(0.5 * log((1 + q_XY) / (1 - q_XY)) + (u_0975) / (sqrt(n-3)) - q_XY /
(2* (n-1)) );
roeInt %roeInt = [0.9018    1.0722]
% => 0.9018 < q_XY < 1.0722 (погр 7 * 10 ^ -2)
```

3. для негруппированных и группированных данных составить уравнения линейной регрессии Y на x и X на y;

```
%%
clc;
mx=0; my=0;Dx=0;Dy=0;Kxy=0
for i = 1:1:n
    mx = mx + X(i); my = my + Y(i);
    Dx = Dx + X(i)^2; Dy = Dy + Y(i)^2;
    Kxy = Kxy + X(i) * Y(i);
end
mx = mx/n; my = my / n;
Dx = Dx/n -mx^2; Dy = Dy/n - my^2;
Kxy = Kxy/n -mx*my; qxy =  Kxy / (sqrt(Dx) * sqrt(Dy));
mess = sprintf('mx = %d\nmy = %d\nDx = %d\nDy = %d\nKxy = %d\nCorrCoef X,Y = %d', mx,my,Dx,Dy,Kxy, qxy);
disp(mess)
messYonX = sprintf('\nВыборочная лин. регрессия Y на x:\nКоэф. перед x = %d\nСвободный член =
%d', Kxy / Dx, my-Kxy*mx/Dx);

messXonY = sprintf('\nВыборочная лин. регрессия X на y:\nКоэф. перед y = %d\nСвободный член =
%d', Kxy / Dy, mx-Kxy*my/Dy);
disp(messYonX); disp(messXonY);
```

```
mx = 1.015600e+01
my = 1.014140e+01
Dx = 3.093517e+01
Dy = 3.699714e+01
Kxy = 3.096939e+01
CorrCoef X,Y = 9.154240e-01

Выборочная лин. регрессия Y на x:
Коэф. перед x = 1.001106e+00
Свободный член = -2.583428e-02

Выборочная лин. регрессия X на y:
Коэф. перед y = 8.370752e-01
Свободный член = 1.666886e+00
```

```
%% Для групп. выборки
clc
n_points = 8;
mx_g = 0; my_g = 0; Dx_g = 0; Dy_g = 0; Kxy_g = 0;
sortedX = sort(X); sortedY = sort(Y);
linSpacedX = linspace(sortedX(1), sortedX(length(sortedX)), n_points);
linSpacedY = linspace(sortedY(1), sortedY(length(sortedY)), n_points);

absFreqX = zeros(1, n_points-1); absFreqY = zeros(1, n_points-1);
for i = 1:1:7
    absFreqX(i) = countInRange(linSpacedX(i), linSpacedX(i+1), sortedX);
    absFreqY(i) = countInRange(linSpacedY(i), linSpacedY(i+1), sortedY);
end
absFreqX; % 6     4     8     9    11     6     6
absFreqY; % 2     4     6     8    14     9     7
mx_g = average(linSpacedX);my_g = average(linSpacedY);
Dx_g = var(linSpacedX, 1); Dy_g = var(linSpacedY, 1);
for i = 1:1:length(linSpacedX)
    Kxy_g=Kxy_g+linSpacedX(i)*linSpacedY(i);
end
Kxy_g = Kxy_g/length(linSpacedX) - mx_g*my_g % 57.9420
qxy_g = Kxy_g / (sqrt(Dx_g) * sqrt(Dy_g)) % 1.0000

mess = sprintf('mx = %d\nmy = %d\nDx = %d\nDy = %d\nKxy = %d\nCorrCoef X,Y = %d',
mx_g,my_g,Dx_g,Dy_g,Kxy_g, qxy_g);
disp(mess)
messYonX = sprintf('\nВыборочная лин. регрессия Y на x:\nКоэф. перед x =
%d\nСвободный член = %d', Kxy_g / Dx_g, my_g-Kxy_g*mx_g/Dx_g);
messXonY = sprintf('\nВыборочная лин. регрессия X на y:\nКоэф. перед y =
%d\nСвободный член = %d', Kxy_g / Dy_g, mx_g-Kxy_g*my_g/Dy_g);
disp(messYonX); disp(messXonY);
```

```
my = 7.890000e+00
Dx = 4.599840e+01
Dy = 7.298679e+01
Kxy = 5.794200e+01
CorrCoef X,Y = 1.000000e+00

Выборочная лин. регрессия Y на x:
Коэф. перед x = 1.259653e+00
Свободный член = -4.467191e+00

Выборочная лин. регрессия X на y:
Коэф. перед y = 7.938697e-01
Свободный член = 3.546368e+00
```

4. для негруппированных данных нанести графики выборочных регрессионных прямых на диаграмму рассеивания.

```
%% 4 для негруппированных данных нанести графики выборочных регрессионных прямых на
диаграмму рассеивания.
clc; syms Y1 x1 X2 y2; %Y = Ax + B; X = Cy + D
A = (Kxy / Dx); B = my-Kxy*mx/Dx; C = Kxy / Dy; D = mx-Kxy*my/Dy;
figure(1); hold on; grid on;title('Регрессионные прямые на графике рассеивания')
GR = scatter(X,Y, 25); %График рассеивания
x1 = linspace(sortedX(1), sortedX(n), 50); %min(X):0.001:max(X);
y2 = -5:0.001:25;
Y1 = A*x1 + B;
X2 = C*y2 + D; %Функции регресс. прямых
plot(Y1, x1, 'm')
plot(y2, X2, 'c')
legend('Точки рассеивания','Y = Ax + B','X = Cy + D')
```


Регрессионные прямые на графике рассеивания
25
20
15
10
5
0
-5
-10
-10
-5
0
5
10
15
20
25
30
Точки рассеивания
Y = Ax + B
X = Cy + D