Theme 8 (part 2). Elements of regression theory
Елементи теорії регресії

A pair linear regression equation

Numerical characteristics:

1) mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$; $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$;

$\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$;

2) variance: $\sigma_x^2 = \overline{x^2} - (\bar{x})^2$; $\sigma_y^2 = \overline{y^2} - (\bar{y})^2$; $\sigma^2_{residuals} = \frac{\sum_i (y_i - \tilde{y}_i)^2}{n-2}$

3) root-mean square deviation: $\sigma_x = \sqrt{\sigma_x^2}$; $\sigma_y = \sqrt{\sigma_y^2}$;

$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}}$.

4) the coefficient of a covariation: $\mu_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$.

5) THE METHOD OF LEAST SQUARES (the system of equations for coefficients b_0, b_1):

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2}; \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}.$$

<u>*Conclusion*</u>: *The regression equation* is $\tilde{y}_x = b_0 + b_1 x$. Coefficient b_1 shows that the increasing *x* by 1 unit gives the increasing (decreasing) *y* by b_1 units.

6) the empirical correlation coefficient (a strength): $r_{xy} = \frac{\mu_{xy}}{\sigma_x \cdot \sigma_y}$

$(-1 \le r_{xy} \le 1)$.

<u>*Conclusion*</u>: If $|r_{xy}| < 0{,}35$ then this correlation is **weak**, if $0{,}35 \le |r_{xy}| \le 0{,}7$ then this correlation is **moderate**, if $|r_{xy}| > 0{,}7$ then this correlation is **strong**.

7) the determination coefficient: $R^2 = r_{xy}^2$.

<u>*Conclusion*</u>: R^2 means that $R^2 \cdot 100\%$ of the total variation in *y* can be explained by the linear relationship between *x* and *y* (as described by the regression equation). The other $\left(1-R^2\right)\cdot 100\%$ of the total variation in *y* remains unexplained.

8) The ***mean error*** $\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \tilde{y}_i}{y_i}\right| \cdot 100\%$ is less than 10% (the allowable limit) then we can use this equation for a forecasting.

9) the elasticity coefficient: $\overline{Э} = b_1 \cdot \frac{\overline{x}}{\overline{y}}$. <u>*Conclusion*</u>: The elasticity coefficient is a number that indicates the percentage change that will occur in one variable (*y*) when the variable *x* changes one percent.

10) ***F-test (Fisher-test)***: H_0 is a statistical nonsignificance of an equation and an indicator of a correlation strength. We compare $F_{emp} = \frac{r_{xy}^2}{1-r_{xy}^2}(n-2)$ with $F_{tabl}(1; n-2)$. <u>*Conclusion*</u>: H_0 is accepted with 95% if $F_{tabl} > F_{emp}$. H_0 is rejected with 95% if $F_{tabl} < F_{emp}$.

11) t-*test (Student test)* for a verification of a significance of a regression equation and a correlation. The assumption H_0 about a random character of model parameters or a correlation coefficient. We compare $t_{b_i} = \frac{b_i}{m_{b_i}}$ and $t_r = \frac{r}{m_r}$ with $t_{0,05}(n-2)$, where $m_{b_0} = \frac{\sigma_{residuals}}{\sigma_x\sqrt{n}}$, $m_{b_1} = \sigma_{residuals}\frac{1}{n\sigma_x}\sqrt{\sum_i x_i^2}$, $m_r = \sqrt{\frac{1-r_{xy}^2}{n-2}}$.

<u>*Conclusion*</u>: H_0 is accepted with 95% if $t_{tabl} > t_{emp}$. H_0 is rejected with 95% if $t_{tabl} < t_{emp}$.

TASKS.

The dependence between the variables x and y was obtained with the help of the experiment and presented by the table:

1) Calculate coefficients of a pair liner regression;
2) Construct the theoretical regression line Y upon X;
3) Calculate the correlation coefficient;
4) Calculate the determination coefficient;
5) Calculate the standard error of a regression;
6) Calculate standard errors of coefficients of a regression;
7) Estimate a statistical significance of coefficients;
8) Estimate a quality of the constructed regression line;
9) Make an analysis of the constructed models.

Task 1. The results of observations of variables X and Y are given as:

a)

X	14	16	18	20	22
Y	15	18	17	19	24

b)

X	1,5	3,0	4,5	6,0	7,5	9,0	10,5
Y	14,0	24,0	33,0	37,0	46,0	51,0	63,0

c)

X	0,25	2,25	4,25	6,25	8,25	10,25	12,25
Y	6,0	9,5	16,0	18,0	23,5	27,0	33,0

d)

X	2,0	3,5	4,0	5,5	7,0	8,5	9,0
Y	10	21	25	33	41	51	63

e)

X	0,5	3,5	6,5	9,5	12,5	15,5	18,5
Y	52,3	48,2	43,5	40,2	38,3	35,2	30,1

f)

X	3,25	4,25	9,25	12,25	13,25	18,25	21,25
Y	22,0	20,0	17,0	15,5	12,5	10,2	8,8

g)

X	17,5	18,5	19,5	20,5	21,5	22,5	23,5	24,5	25,5
Y	25,56	27,68	30,70	33,20	38,76	40,52	43,32	50,00	52,00

TASK 2.

X	100	150	200	250	300
Y	60	35	20	20	15

TASK 3.

№	45,1	68,8
1	59,0	61,2
2	57,2	59,9
3	61,8	56,7
4	58,8	55,0
5	47,2	54,3
6	55,2	49,3
7	45,1	68,8

Task 4. $\sum x_i = 1027,\ n = 12,\ \sum y_i = 1869,\ \sum x_i y_i = 161808,\ \sum x_i^2 = 89907,$

$\sum y_i^2 = 294377,\ \sum_{i=1}^{n} \left| \frac{y_i - \tilde{y}_i}{y_i} \right| = 0{,}688,\ \sum_i (y_i - \tilde{y}_i)^2 = 1600.$

Task 5.

$\sum x_i = 384{,}3,\ n = 7,\ \sum y_i = 405{,}2,\ \sum x_i y_i = 22162{,}34,\ \sum x_i^2 = 21338{,}41,$

$\sum y_i^2 = 23685{,}76,\ \sum_{i=1}^{n} \left| \frac{y_i - \tilde{y}_i}{y_i} \right| = 0{,}57,\ \sum_i (y_i - \tilde{y}_i)^2 = 200{,}56.$

Task 6. $\sum x_i = 8{,}237,\ n = 9,\ \sum y_i = 3{,}931,\ \sum x_i y_i = 4{,}2087,\ \sum x_i^2 = 9{,}2334,$

$\sigma_y = 0{,}08,\ \sum_i (y_i - \tilde{y}_i)^2 = 0{,}0014.$

Task 7. The results of 15 observations of variables X and Y are given as:

a) $\sum_{i=1}^{15} x_i = 28,91;\ \sum_{i=1}^{15} x_i^2 = 57,061;\ \sum_{i=1}^{15} y_i = 121,6;\ \sum_{i=1}^{15} y_i^2 = 1017,268;$

$\sum_{i=1}^{15} x_i y_i = 228,05.$

b) $\sum x_i = 26,61;\ \sum y_i = 131,95;\ \sum x_i y_i = 228,272;\ \sum x_i^2 = 48,319;$

$\sum y_i^2 = 1194,959.$

c) $\sum x_i = 23,19;\ \sum y_i = 149,95;\ \sum x_i y_i = 227,238;$

$\sum x_i^2 = 36,829;\ \sum y_i^2 = 1523,986.$

You need

1) to calculate coefficients b_0 and b_1 of a linear regression equation and standard errors of coefficients;
2) to construct confidence intervals of coefficients for 90 % and 99 %;
3) to calculate the determination coefficient R^2;
4) make an analysis of the constructed models.

Task 8. The results of observations of variables X and Y are given as:

A number of employees	Average productiv-ity, X, $	Average pay, Y,$	Standard deviation of pay, $
4	9 320	3 320	740
9	8 630	3 640	850
18	8 050	3 900	730
48	9 320	4 120	820
89	8 600	4 090	950
159	9 120	4 200	1 100
319	9 540	4 380	1 250
899	9 730	4 500	1 290
1 569	10 120	4 610	1 350
3 559	10 740	4 800	1 100
6 000	11 200	5 000	1 520

Construct the regression equation $Y = b_0 + b_1 X + e$ using the method of least squares.

STUDENT'S DISTRIBUTION

The significance level α				
The degree of freedom k	$\alpha = 0,01$	$\alpha = 0,02$	$\alpha = 0,05$	$\alpha = 0,1$
1	63,7	31,82	12,7	6,31
2	9,92	6,97	4,30	2,92
3	5,84	4,54	3,18	2,35
4	4,60	3,75	2,78	2,13
5	4,03	3,37	2,57	2,01
6	3,71	3,14	2,45	1,94
7	3,50	3,00	2,36	1,89
8	3,36	2,90	2,31	1,86
9	3,25	2,82	2,26	1,83
10	3,17	2,76	2,23	1,81
11	3,11	2,72	2,20	1,80
12	3,05	2,68	2,18	1,78
13	3,01	2,65	2,16	1,77
14	2,98	2,62	2,14	1,76
15	2,95	2,60	2,13	1,75

FISHER'S DISTRIBUTION for 0.05

k_2	k_1											
	1	2	3	4	5	6	7	8	9	10	11	12
1	161¤	200¤	216¤	225¤	230¤	234¤	237¤	239¤	241¤	242¤	243¤	244¤
2	18,5¤	19,0¤	19,2¤	19,2¤	19,3¤	19,3¤	19,3¤	19,4¤	19,4¤	19,4¤	19,4¤	19,4¤
3	10,1¤	9,55¤	9,28¤	9,13¤	9,01¤	8,94¤	8,88¤	8,84¤	8,81¤	8,78¤	8,76¤	8,74¤
4	7,71¤	6,94¤	6,59¤	6,39¤	6,26¤	6,16¤	6,09¤	6,04¤	6,00¤	5,96¤	5,93¤	5,91¤
5	6,61¤	5,79¤	5,41¤	5,19¤	5,05¤	4,95¤	4,88¤	4,82¤	4,78¤	4,74¤	4,70¤	4,68¤
6	5,99¤	5,14¤	4,76¤	4,53¤	4,39¤	4,28¤	4,21¤	4,15¤	4,10¤	4,06¤	4,03¤	4,00¤
7	5,59¤	4,74¤	4,35¤	4,12¤	3,97¤	3,87¤	3,79¤	3,73¤	3,68¤	3,63¤	3,60¤	3,57¤
8	5,32¤	4,46¤	4,07¤	3,84¤	3,69¤	3,58¤	3,50¤	3,44¤	3,39¤	3,34¤	3,31¤	3,28¤
9	5,12¤	4,26¤	3,86¤	3,63¤	3,48¤	3,37¤	3,29¤	3,23¤	3,18¤	3,13¤	3,10¤	3,07¤
10	4,96¤	4,10¤	3,71¤	3,48¤	3,33¤	3,22¤	3,14¤	3,07¤	3,02¤	2,97¤	2,94¤	2,91¤
11	4,84¤	3,98¤	3,59¤	3,36¤	3,20¤	3,09¤	3,01¤	2,95¤	2,90¤	2,86¤	2,82¤	2,79¤
12	4,75¤	3,88¤	3,49¤	3,26¤	3,11¤	3,00¤	2,92¤	2,85¤	2,80¤	2,76¤	2,72¤	2,69¤
13	4,67¤	3,80¤	3,41¤	3,18¤	3,02¤	2,92¤	2,84¤	2,77¤	2,72¤	2,67¤	2,63¤	2,60¤
14	4,60¤	3,74¤	3,34¤	3,11¤	2,96¤	2,85¤	2,77¤	2,70¤	2,65¤	2,60¤	2,56¤	2,53¤
15	4,54¤	3,68¤	3,29¤	3,06¤	2,90¤	2,79¤	2,70¤	2,64¤	2,59¤	2,55¤	2,51¤	2,48¤