Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 37(3), 2010, Pages 101–109
ISSN: 1223-6934

Correlation Interpretation on Diabetes Mellitus Patients

Dana Dănciulescu

Abstract. The correlation is a causal, complementary, parallel, or reciprocal relationship, especially a structural, functional, or qualitative correspondence between two comparable entities. This article offers an example on how to compute and interpret the correlation coefficient. The parameters used are three main features (age, sex and residence) of patients with diabetes mellitus.

2010 Mathematics Subject Classification. Primary: 62M10; Secondary: 62J05, 62M99.
Key words and phrases. correlation coefficient, regression, diabetes mellitus.

1. Introduction

It is known that socio-economic phenomena variability is mostly caused by simultaneous action of several factors; some of these factors favor the evolution of a phenomenon, others hinder or even work in reverse.

Correlation is used to quantify and analyze causal relationship between socio-economic phenomena. Specific statistical methods offer knowledge, mainly of the following: [6]:

- existence of causal relationships between phenomena;
- each factor contribution to overall variability phenomena effect;
- intensity of causal links between socio-economic phenomena and processes;
- evolutionary trends of correlation between events.

The correlation can be defined as a causal, complementary, parallel, or reciprocal relationship, especially a structural, functional, or qualitative correspondence between two comparable entities. The correlative tuple contains two or more variables from which one is called effect variable, while the others are cause variables. The effect variable is in fact a result variable while the cause variables are factorial ones.

The correlation can be classified depending on [12]:

(1) the number of variables from the correlative tuple
   (a) simple correlation;
   (b) multiple correlation;
(2) the factorial links
   (a) direct correlation;
   (b) inverse correlation;
(3) the causal links
   (a) linear correlation;
   (b) non-linear correlation.

This article represents an example on how to compute and interpret the correlation. The statistic data used is a sample of 80 patients with diabetes mellitus. The parameters taken in consideration are: age, sex and residence (urban/rural).

Received July 05, 2010. Revision received September 01, 2010.

2. State of Art

Correlation coefficient is a quantitative value between -1 and 1 and it describes the relationship between two or more variables. The extreme values indicate a perfect relationship between variables while 0 shows a total lack of linear relationship. To obtain a proper interpretation of values the results must be compared with certain defaults on tables of correlations depending on the number of subjects, type of connection desired and materiality [13].

The most common types of parametric correlations are [11]:

(1) Simple correlation coefficient (Bravais-Pearson's)
(2) Eneahoric correlation coefficient;
(3) Coefficient of partial correlation;
(4) Biserial and triserial correlation coefficients.

In this article we use the statistical-mathematical function in the study of simple regression and also the simple correlation intensity quantification methodology.

To analyze the correlation we must take in consideration two important factors [10]:

- regression – defined as the relationship between the mean value of a random variable and the corresponding values of one or more independent variables. It helps us to determine the determining factors contributions to variability effect phenomena;
- correlation intensity – which is summarized by correlation coefficients.

The linear regression of order 1 is expressed by [8, 3]:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

y = dependent variable
x = independent variable
β_0 = the initial value of y
β_1 = the modification of y caused by the changes of x
ϵ = variable's error

In order to compute the parameters β_0 and β_1, we use estimator parameters of the following linear function:

$$\widehat{y} = b_0 + b_1 x$$

where

$$b_1 = \frac{cov(x, y)}{s_x^2}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

The elements for computing these two parameters are determined using the following expressions:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

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

However, it can often happen that empirical values y_i may differ from the theoretical ones $\overline{y}$; these elements represent residual values and are caused by random factors. The residual value are statistically known as error [4].

To compute the average error we use the following expression:

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

where SSE is the sum of squared errors of estimation given by:

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_1})^2$$

Verifying the null hypothesis will determine if b_1 estimates correctly β_1:

$$H_0 : \beta_1 = 0$$

Its alternative is represented by:

$$H : \beta_1 \neq 0$$

In order to verify the null hypothesis we use the test t given by the relation:

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

where

$$S_{b_1} = \frac{S_e}{\sqrt{(n-1)s_x^2}}$$

If the variable's errors follow the normal distribution law, the statistic test t used must meet $n-2$ degrees of freedom [5].

The null hypothesis is rejected if:

$$t > t_{\alpha/2,n-2} \quad or \quad t < -t_{\alpha/2,n-2}$$

The determining coefficient is used to determine the contribution of one of the factors which influences the variability of the dependent phenomenon:

$$R^2 = \frac{[cov(x,y)]^2}{s_x^2 s_y^2}$$

This coefficient R^2 shows the proportion of the dependent variable's variation y determined by the influence of the independent variable's variation x.

Correlation coefficient r shows the intensity of the causal link between the two variables. Both coefficients are based on Pearson's relationship:

$$r = \sqrt{R^2} = \frac{cov(x,y)}{s_x s_y}$$

The meaning of the correlation coefficient r is emphasized by values between -1 and 1:

- values between 0 and 1 show a direct correlation of increasingly intense as they approach one;
- values between 0 and -1 show an inverse correlation of increasingly intense as close to -1;
- zero value points out that between the two variables there is no connection.

In practice the interval between -1 and 1 is generally refined as follows [14]:

- If $0 \leq r < 0.2$ there is no significant relation between variables;
- If $0.20 \leq r \leq 0.50$ the relationship between variables is low;
- If $0.50 \leq r < 0.75$ the relationship between variables is average;

- If $0.75 \leq r < 0.95$ the relationship between variables is strong;
- If $0.95 \leq r \leq 1$ there is a functional relation between the two variables.

The correlation between two parameters can be computed using:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{1}$$

where x_i and y_i are the two parameters.

There are situations when the correlation between two variables can be difficult to interpret because of the existence of a third variable which influences the dependence between the other two. In such cases there must be used the formula for the partial correlation coefficient given by:

$$r_{12.3} = \frac{r_{12} - r_{13} * r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

where r_{12}, r_{23} and r_{13} are the correlation coefficients.

3. Correlation data on diabetes mellitus

The sample considered in this article consists in 80 patients with diabetes mellitus, a disease known as mostly caused by genetic inheritance. The patients are all located in the same district and they where registered with this disease in January, February or March 2010. The list of the patients with their characteristics is presented in the table 1. The last column of the table contains the squares of the age values which are needed in the equation 1.

To apply the formulas presented in the section above we replace the sex and the residence with numeric values as follows:

$$\begin{array}{ll} \text{F} = 1 & \text{U} = 1 \\ \text{M} = 2 & \text{R} = 2 \end{array}$$

We first compute the correlation between age and sex. To obtain r with formula 1 we need:

$$\sum_{i=1}^{80} x_i y_i = 6843$$

$$\sum_{i=1}^{80} x_i = 4580$$

$$\sum_{i=1}^{80} y_i = 121$$

$$\sum_{i=1}^{80} x_i^2 = 272180$$

$$\sum_{i=1}^{80} y_i^2 = 203$$

TABLE 1. Patients with diabetes mellitus

Number	Age	Sex	Residence	Age^2
1	47	F	U	2209
2	57	F	U	3249
3	67	F	R	4489
4	48	M	U	2304
5	55	M	U	3025
6	66	M	R	4356
7	56	M	U	3136
8	50	M	R	2500
9	49	F	U	2401
10	56	F	U	3136
11	61	M	U	3721
12	59	M	U	3481
13	40	M	U	1600
14	67	F	R	4489
15	53	F	R	2809
16	47	M	R	2209
17	59	M	U	3481
18	68	M	U	4624
19	54	F	U	2916
20	59	M	U	3481
21	74	M	U	5476
22	55	F	U	3025
23	49	F	U	2401
24	72	F	U	5184
25	80	F	U	6400
26	57	M	U	3249
27	67	F	U	4489
28	65	M	U	4225
29	39	M	U	1521
30	45	F	U	2025
31	60	M	U	3600
32	73	F	U	5329
33	40	M	U	1600
34	66	M	U	4356
35	71	F	U	5041
36	40	M	U	1600
37	69	F	U	4761
38	40	M	U	1600
39	66	M	U	4356
40	72	F	U	5184

So we have:

$$\begin{aligned} r &= \frac{80 * 6843 - 4580 * 121}{\sqrt{80 * 272180 - 4580^2}\sqrt{80 * 203 - 121^2}} \\ &= -0,188684 \end{aligned} \tag{2}$$

Number	Age	Sex	Residence	Age^2
41	54	M	U	2916
42	57	M	U	3249
43	53	F	R	2809
44	61	F	U	3721
45	46	F	U	2116
46	75	M	U	5625
47	44	M	U	1936
48	49	M	U	2401
49	50	M	U	2500
50	55	F	U	3025
51	33	M	U	1089
52	66	M	U	4356
53	53	M	U	2809
54	62	F	R	3844
55	55	M	U	3025
56	55	F	R	3025
57	63	F	R	3969
58	63	F	R	3969
59	43	M	U	1849
60	62	M	U	3844
61	56	M	U	3136
62	57	F	U	3249
63	69	M	R	4761
64	53	F	U	2809
65	63	F	U	3969
66	69	F	R	4761
67	21	F	U	441
68	68	M	R	4624
69	72	F	R	5184
70	72	F	R	5184
71	54	M	R	2916
72	53	M	U	2809
73	61	F	R	3721
74	52	F	U	2704
75	70	F	R	4900
76	38	F	U	1444
77	80	F	R	6400
78	48	F	U	2304
79	50	M	U	2500
80	57	M	R	3249

This value for r shows a lack of a significant relation between age and sex for the patients with diabetes mellitus. So we cannot say that women (or men) with the age between a certain interval are predisposed to this disease. To see if there is a correlation between age and residence we need:

$$\sum_{i=1}^{80} x_i y_i = 5898$$

$$\sum_{i=1}^{80} x_i = 4580$$

$$\sum_{i=1}^{80} y_i = 101$$

$$\sum_{i=1}^{80} x_i^2 = 272180$$

$$\sum_{i=1}^{80} y_i^2 = 143$$

So the correlation is:

$$\begin{aligned} r &= \frac{80 * 5898 - 4580 * 101}{\sqrt{80 * 272180 - 4580^2}\sqrt{80 * 143 - 101^2}} \\ &= 0,294492 \end{aligned} \tag{3}$$

Being in the interval $[0.2, 0.5]$, r is emphasizing a low relationship between age and residence. As we can see from the table 1 the average age of the patients from the country side is 62, while the average age from the others is 55.55. Also, we can see that the number of urban patients is almost three times higher than the number of the ones from the country side. The correlation between sex and residence is:

$$\sum_{i=1}^{80} x_i y_i = 149$$

$$\sum_{i=1}^{80} x_i = 121$$

$$\sum_{i=1}^{80} y_i = 101$$

$$\sum_{i=1}^{80} x_i^2 = 203$$

$$\sum_{i=1}^{80} y_i^2 = 143$$

$$\begin{aligned} r &= \frac{80 * 149 - 121 * 101}{\sqrt{80 * 203 - 121^2}\sqrt{80 * 143 - 101^2}} \\ &= -0,213849 \end{aligned} \tag{4}$$

The correlation slightly exceeds the threshold of -0.2 showing that it may be a very low relationship between these two parameters. The minus sign shows an inverse direction of the correlation. However, the value cannot be interpreted as a sure relationship. In such cases, it is recommended to extend the sample to a larger number of patients and repeat all the calculations. When the sample is much higher the patients can

be grouped depending on their age. The intervals length is determined with Sturges' formula [7]:

$$l = \frac{x_{max} - x_{min}}{1 + 3.332 \lg_n}$$

where x_{max} is the highest age from the sample and x_{min} is the minimum. In our case the length is:

$$l = \frac{80 - 21}{1 + 3.32 \lg 80} = \frac{59}{7.31} = 8.07$$

So, the age intervals will be:

$$[21, 29], (29, 37], (37, 45], (45, 53],$$

$$(53, 61], (61, 69], (69, 77], (77, 80]$$

Then, the data from each interval is numbered and then the correlation is computed using the same formulas. The results will be largely the same but this artifice greatly facilitates the calculations when we are dealing with a large sample.

4. Conclusions

The correlation coefficient shows if there is a connection between two parameters and how strong this connection is. All the three results obtained have low values emphasizing a low relationship or even a total absence of it. If we would have chosen parameters like glucose level, polydipsia or polyuria the values of the correlation coefficient would have been much higher since we all know that these factors are some of the most important ones for this disease [1, 2]. We chose these three parameters since they are often confused with factors that influence the diabetes mellitus. From the sample we can see that the number of patients from the country side is much lower and also these patients have an older age. The difference between the average age is not so great – but it is significance – but the difference between the minimum values and the maximum ones is extreme (minimum from rural is 47 while for urban is 21, maximum from rural is 80 while for urban is 74).

References

[1] Diabetes Mellitus, WHO Technical Report Series, 1985.
[2] *Clinical Practice Guidelines for Treatment of Diabetes Mellitus*, Expert Committee of the Canadian Diabetes Advisory Board, Canadian Medical Association Journal **147** (1992), no. 5.
[3] E. Achtert, C. Bohm, H.P. Kriegel, P. Kroger and A. Zimek, On exploring complex relationships of correlation clusters., *In Proc. SSDBM* (2007).
[4] J. Bartko, On various intraclass correlation reliability coefficients, *Psychological Bulletin* (1976), 762–765.
[5] L. Crocker and J. Algina, Introduction to classical and modern test theory., *Fort Worth: Holt*, Rinehart and Winston, (1986).
[6] A.L. Edwards, The Correlation Coefficient, *W. H. Freeman (ed) An Introduction to Linear Regression and Correlation*, San Francisco, CA, (1976), 33–46.
[7] L.A. Goodman and W.H. Kruskal, Measures of association for cross-classifications III: Approximate sampling theory, *J. Amer. Statistical Assoc.* **58** (1963), 310–364.
[8] M.G. Kendall, *Rank Correlation Methods*, Charles Griffin & Co., 1955.
[9] J.F. Kenney and E.S. Keeping, Linear Regression and Correlation, *Mathematics of Statistics (3rd ed. Princeton)*, NJ: Van Nostrand, (1962).
[10] E.L. Lehmann, *Testing statistical hypotheses. (2nd ed.)*, New York: Wiley, 1986.
[11] J.L. Rodgers and W.A. Nicewander, Thirteen ways to look at the correlation coefficient, *The American Statistician* **42** (1988), no. 1, 59-66.

[12] G.W. Snedecor and W.G. Cochran, The Sample Correlation Coefficient r and Properties of r, *Statistical Methods, (7th ed. Ames)*, IA: Iowa State Press, (1980).
[13] M.R. Spiegel, Correlation Theory, *Theory and Problems of Probability and Statistics (2nd ed.)*, New York: McGraw-Hill, (1992).
[14] B.J. Winer, *Statistical principles in experimental design (2nd ed.)*, New York: McGraw-Hill, 1971.

(Dana Dănciulescu) Department of Economics, University of Craiova, Al.I. Cuza Street, No. 13, Craiova RO-200585, Romania, Tel. & Fax: 40-251412673
E-mail address: danadanciulescu@central.ucv.ro