Review of: "Decoding the Correlation Coefficient: A Window into Association, Fit, and Prediction in Linear Bivariate Relationships"

Karoly Heberger[1]

1 Research Centre for Natural Sciences

Potential competing interests: No potential competing interests to declare.

Title: Decoding the Correlation Coefficient: A Window into Association, Fit, and Prediction in Linear Bivariate Relationship,s preprint v2

Qeios ID: BNOKLP

Author: S A Hamed Hosseini

Open Access

https://doi.org/10.32388/O1ZP41

The title promises a lot, but the text only reveals that the author addresses a partial question related to the correlation coefficient, and even that only superficially. I always admire the courage of those who, having so little knowledge and insight, set out to work on such a multifaceted problem.

The first is the dimensional problem. Whereas (r) is a scale free, the slope has dimension, for example if y is concentration and x is time, measured in mol/l and s respectively, then the dimension of the slope is concentration divided by time measured in the example mol/l/s.

Hence, to find the relation between r and the slope is of course case dependent and no general conclusion can be drawn. Naturally, both variables can be standardized, However, the correlation coefficient of standardized variables are called congruence- (or cosine) coefficient, which are not even mentioned in this text. The usage of intercept (a) contradict this assumption.

Secondly, the author deals with a sub-topic of linear regression when both variables are subject to errors. It is a well-know problem and slowly dying out. A recent review summarizes the main knowledge in this field [1]. The basic formula (Eq. (9) in ref. [1]) is missing:

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y \rightarrow \frac{Cov(X,Y)}{Var(X)} = \frac{\beta\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \qquad (9)$$

The author does not even reveal the main feature of orthogonal regression: if both variables are subject to error; then, it is no use (or it is impossible) to define dependent and independent (predictor) variables.

Much less known is that “if x is also subject to random errors of measurement, linear regression analysis yields an underestimate of the slope and a correspondingly biased estimate of the intercept.” [2].

Thirdly: If “coefficient of determination" is confused with "correlation coefficient"; then the author has no models in mind. Fortunately, a kind reviewer (Szabolcs Blazsek) defines the model properly. What is missing here: a correlation coefficient can also be defined between measured and predicted y values (it is called determination coefficient) and it can be done even in different ways, which can lead to different results. A detailed discussion of this issue can be found in ref. [3].

It should also be mentioned that the term correlation coefficient is not ONE single quantity. The most frequently used one is the Pearson product moment correlation coefficient, but rank correlation coefficients (Spearman’s rho and Kendall Tau (Gamma)) are also available and they can be more useful in certain circumstances. Binary version of correlation coefficient is also known as Pearson’s phi, or Matthews's correlation coefficient.

The hardly understandable statement “… cannot assert with certainty that a higher correlation coefficient (r) never implies a stronger correspondence in change." First of all, it is a very weak statement, secondly it does not reveal the degree of freedom. Namely, significance of correlation coefficient depends largely on the degree of freedom. C.f. Bevigton’s table [4].

[1] Pallavi, Sandeep Joshi, Dilbag Singh, Manjit Kaur, Heung-No Lee: Comprehensive Review of Orthogonal Regression and Its Applications in Different Domains, Archives of Computational Methods in Engineering 29 (2022)4027-4047.

https://doi.org/10.1007/s11831-022-09728-5

[2] Louis Meites, H. C. Smit, and G. Kateman: The Effects of Errors in Measuring the Independent Variable in Least-Squares Regression Analysis, Analytica Chimica Acta, 164 (1984) 287-291. https://doi.org/10.1016/S0003-2670(00)85642-1

[3] K. Heberger: Discrimination between Linear and Non-Linear Models Describing Retention Data of Alkylbenzenes in Gas-Chromatography Chromatographia, 29(7/8) (1990) 375-383. https://doi.org/10.1007/BF02261306

[4] P.R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, Table C-3 in the appendix. McGraw-Hill Book Co., New York, 1969.

Appendix:

P.R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill Book Co., New York, 1969.

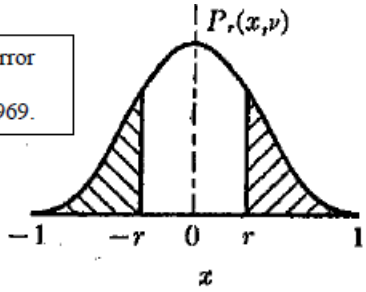

Table C-3 Linear-correlation coefficient. The linear-correlation coefficient r vs. the number of observations N and the corresponding probability $P_e(r,N)$ of exceeding r in a random sample of observations taken from an uncorrelated parent population ($\rho = 0$)

N \ P	0.50	0.20	0.10	0.050	0.020	0.010	0.005	0.002	0.001
3	0.707	0.951	0.988	0.997	1.000	1.000	1.000	1.000	1.000
4	0.500	0.800	0.900	0.950	0.980	0.990	0.995	0.998	0.999
5	0.404	0.687	0.805	0.878	0.934	0.959	0.974	0.986	0.991
6	0.347	0.608	0.729	0.811	0.882	0.917	0.942	0.963	0.974
7	0.309	0.551	0.669	0.754	0.833	0.875	0.906	0.935	0.951
8	0.281	0.507	0.621	0.707	0.789	0.834	0.870	0.905	0.925
9	0.260	0.472	0.582	0.666	0.750	0.798	0.836	0.875	0.898
10	0.242	0.443	0.549	0.632	0.715	0.765	0.805	0.847	0.872
11	0.228	0.419	0.521	0.602	0.685	0.735	0.776	0.820	0.847
12	0.216	0.398	0.497	0.576	0.658	0.708	0.750	0.795	0.823
13	0.206	0.380	0.476	0.553	0.634	0.684	0.726	0.772	0.801
14	0.197	0.365	0.458	0.532	0.612	0.661	0.703	0.750	0.780
15	0.189	0.351	0.441	0.514	0.592	0.641	0.683	0.730	0.760
16	0.182	0.338	0.426	0.497	0.574	0.623	0.664	0.711	0.742
17	0.176	0.327	0.412	0.482	0.558	0.606	0.647	0.694	0.725
18	0.170	0.317	0.400	0.468	0.543	0.590	0.631	0.678	0.708
19	0.165	0.308	0.389	0.456	0.529	0.575	0.616	0.662	0.693
20	0.160	0.299	0.378	0.444	0.516	0.561	0.602	0.648	0.679
22	0.152	0.284	0.360	0.423	0.492	0.537	0.576	0.622	0.652
24	0.145	0.271	0.344	0.404	0.472	0.515	0.554	0.599	0.629
26	0.138	0.260	0.330	0.388	0.453	0.496	0.534	0.578	0.607
28	0.133	0.250	0.317	0.374	0.437	0.479	0.515	0.559	0.588
30	0.128	0.241	0.306	0.361	0.423	0.463	0.499	0.541	0.570
32	0.124	0.233	0.296	0.349	0.409	0.449	0.484	0.526	0.554
34	0.120	0.225	0.287	0.339	0.397	0.436	0.470	0.511	0.539
36	0.116	0.219	0.279	0.329	0.386	0.424	0.458	0.498	0.525
38	0.113	0.213	0.271	0.320	0.376	0.413	0.446	0.486	0.513
40	0.110	0.207	0.264	0.312	0.367	0.403	0.435	0.474	0.501
42	0.107	0.202	0.257	0.304	0.358	0.393	0.425	0.463	0.490
44	0.104	0.197	0.251	0.297	0.350	0.384	0.416	0.453	0.479
46	0.102	0.192	0.246	0.291	0.342	0.376	0.407	0.444	0.469
48	0.100	0.188	0.240	0.285	0.335	0.368	0.399	0.435	0.460
50	0.098	0.184	0.235	0.279	0.328	0.361	0.391	0.427	0.451
60	0.089	0.168	0.214	0.254	0.300	0.330	0.358	0.391	0.414
70	0.082	0.155	0.198	0.235	0.278	0.306	0.332	0.363	0.385
80	0.077	0.145	0.185	0.220	0.260	0.286	0.311	0.340	0.361
90	0.072	0.136	0.174	0.207	0.245	0.270	0.293	0.322	0.341
100	0.068	0.129	0.165	0.197	0.232	0.256	0.279	0.305	0.324

July 15 / 2023 referee: