Improving Linearity in Health Science Investigations

Satyendra Nath Chakrabartty*

Indian Statistical Institute, Indian Institute of Social Welfare and Business Management, Indian Ports Association

***Corresponding author:**
Satyendra Nath Chakrabartty

✉ Chakrabarttysatyendra139@gmail.com

Tel: 919831597909

Indian Statistical Institute, Indian Institute of Social Welfare and Business Management, Indian Ports Association

Citation: Chakrabartty SN (2023) Improving Linearity in Health Science Investigations. Health Sci J. Vol. 17 No. 4: 1010.

Abstract

Correlation and linear regression are frequently used to evaluate the degree of linear association between two variables and also to find the empirical relationship. However, violations of assumptions often give results which are not valid. High value of correlation coefficient is taken as degree of linearity between two variables and attempt is made to fit linear regression equation. However, linearity implies high correlation but the converse is not true. The paper describes with examples that concept of linearity is different from correlations, effect of violation of assumptions of correlations and linear regressions and suggests procedures to improve correlation between two variables which can be extended to multi variables.

Keywords: Linearity; Correlation coefficient; Standard error; Normal distribution; Generalized inverse

Received: 28-Mar-2023, Manuscript No. Iphsj-23-13689; **Editor assigned:** 30-Mar-2023, Pre-QC No. Iphsj-23-13689 (PQ); **Reviewed:** 13-April-2023, QC No. Iphsj-23-13689 **Revised:** 18-April-2023, Manuscript No. Iphsj-23-13689 (R); **Published:** 25-April-2023, DOI: 10.36648/1791-809X.17.4.1010

Introduction

Correlations are often used in various fields of research. There are different kinds of correlations depending on nature of variables. Cause and effect relationship along with direction of the linear relationship between two variables, is reflected by Pearsonoian correlation, assumptions of which include: measurement on each variable is at least interval level, data on each variable follows normal distribution and has no outliers, etc. However, correlation does not always imply causation [1]. Correlation between two variables could be due to a third variable affecting both the variables under study viz. Item reliability in terms of item-total correlation. Variables may be correlated over time where data is longitudinal. For example, earth's temperature and levels of greenhouse gases are positively correlated. Estimating correlation between two such trending variables after removing the trend is desirable [2].

By definition, correlation between X and Y is the ratio of Cov(X,Y) and product of SD(X)and SD(Y). Thus, average of k-number of correlations r̄=(∑_(j=1)^k raja)/k is meaningless for correlations with mixed signs and of same sample size (Field, 2003). However, computation of average inter-item correlations is used in psychological literature to reflect level of consistency of a test and is regarded as a quality of test as a whole. Correlation between two variables (r_XY) is high if the ratio of change in one variable (Y) due to unit change in the second variable(X) is constant for all values of X [3].

Interpretation and use of correlation is important for measurement by practitioners and researchers since simple correlations are used in various studies, including multivariate statistical procedures such as multiple regressions, ANOVA, principal component analysis (PCA), factor analysis (FA), path analysis, structural equation modeling, etc., each of which uses simple correlations and/or their extensions. Interpretation of correlation as proportion of pairs with identical values of the two variables or as the probability of correct prediction of one of the variable with knowledge of the other is wrong [4, 5]. A popular way of interpretation of correlation is to indicate the extent by which variance in one variable is explained by the second variable by computing 〖r_XY〗^2 known as coefficient of determination. For example, 〖r_XY〗^2=0.64 suggests that X accounts for 64% of the variance of Y. The shared variance between X and Y is a key concept for statistics with multiple predictor variables (e.g., factorial ANOVA, multiple regression) and is a common measure of effect size (R^2 and η^2). Rodgers and Nice wander (1988) described 13 ways of interpreting a correlation. Another interpretation of correlation as the proportion of matches was suggested [4].

Correlation coefficient r_XY∈[-1,1] is taken as degree of linearity between two variables. If correlation between X and Y (r_XY) is high i.e.|r_XY |≈1, the variables are usually taken as linearly related and attempt is made to establish linear regression of the form Y=α_1+β_1 X+ ϵ_YX or X=α_2+β_2 Y+ϵ_XY. The assumption of normal distribution is not needed to estimate the regression coefficients (β's) but ϵ_YX and ϵ_XY must be normally distributed with mean = 0 and constant variance (homoscedasticity).

However, Loco, et al. (2002) found that the investigated curves were characterized by a high correlation coefficient (r > 0.997) but the straight-line model was rejected at the 95% confidence level on the basis of the Lack-of-fit and Mandel's fitting test.

The paper describes with examples that concept of linearity is different from correlations, effect of violation of assumptions of correlations and linear regressions and suggests procedures to improve linearity between two variables which can be extended to multi variables (**Table 1**).

Correlation and Linearity

Observations:

If X is increasing and Y is decreasing, r_XY is negative viz. r_(X,1/X) , r_(X,CosX), etc.

r_(X,f(X)) |≥0.92 for non-linear f(X) = X^2,X^3, 〖log〗_10^X, Cos X and Sin X despite non-linear relationship between X and f(X)

Maximum improvement in correlation was observed for X and f(X) = Sin X where r_(X,SinX)=0.99982 followed by r_(X,CosX)=(-) 0.97156. Thus, trigonometric transformations like f(X) = Sin X or f(X) = Cos X tend to improve absolute magnitude of correlation coefficient.

Correlation may not always imply linearity. Scatter plot may throw more light on linearity and validity of linear regression line.

The above can be summarized as "Linearity implies high correlation but the converse is not true".

Question therefore arises on how to know linearity between two variables. A simple way to check linearity between Y and X is to see whether (Resulting change in Y)/(Unit change in X) is constant for all values of X. In other words, one may check for constant slope of the straight line connecting X and Y by considering (Y_i-Y_(i+1))/(X_i-X_(i+1)) and checking whether the ratio is constant for all values of i. If yes, (Y_i-Y_(i+1))/(X_i-X_(i+1)) can be taken as slope of the straight line (β). Checking of (Y_i-Y_(i+1))/(X_i-X_(i+1)) for few illustrative non-linear functions of X are shown in Table 2.

In fact, absolute value of r_XY could be high i.e. |r_XY |≈1 even if X and Y are related by a non-linear fashion. For example, if X takes integer values from 1, 2, 3… 30, correlation between X and several non-linear function of X are high as shown below (**Table 2**).

Visual approaches to check normality include among others the output of a quantile–quantile or Q–Q plot which is a scatter-plot of the quantiles of a theoretical normal data set (on X-axis) and the quantile of the actual sample data set (on Y-axis). If the data are normally distributed, the data points on the Q–Q plot will be closely aligned with the straight line with slope 1. If the individual data points are away from the diagonal line, data are not normally distributed. Alternatively, linearity can be tested by first fitting a linear regression line of the form (say) Y= α+βX+ϵ followed by finding predicted values of Y as Y̑ and then testing significance of variance of error scores S_E^2 or standard error S_E=√(1/n ∑〖(Y_i-Y ̂_i)〗^2) = S_Y √(1-r^2) where n denotes number of observations and E=(Y-Y̑). Note that higher absolute value of correlation will result in lower value of S_E and may lead to acceptance of H_0: S_E^2=0. Normal probability plot of error score for X and illustrative non-liner function of X are shown below:

Normal probability of error score for predicting X^2 on X is given in (**Figure 1**).

Error score for predicting X^2 on X did not pass the normality test. AD statistic was 1.048 and p-value 0.007955658 (**Figures 2 & 3**)

For better visualization of linearity one can draw a scatter plot

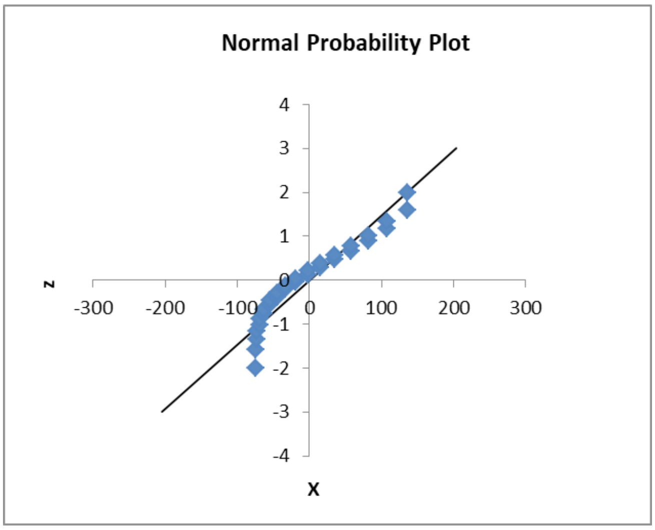

Figure 1 Normal probability plot of error score for X^2= α+βX.

Table 1. Correlation between X and Non-Linear function of X.

	X	X^2	$^1/_X$	X^3	$\log_{10}^X$	Cos X	Sin X
X	1	0.97029	-0.64789	0.92011	0.92064	-0.97156	0.99982
X^2		1	-0.50445	0.98629	0.81179	-0.99998	0.96559
$^1/_X$			1	-0.42219	- 0.87699	0.50689	-0.65623
X^3				1	0.72716	-0.98529	0.91251
Log_{10}^X					1	-0.81425	0.92630
Cos X						1	-0.96696
Sin X							1

Table 2. Checking of (Y_I-Y_(I+1))/(X_I-X_(I+1))=K With different Y=F(X).

$1 \le X \le 30$ (X Is +Ve Integer)		$Y = X^2$		$\frac{Y_I - Y_{I+1}}{X_I - X_{I+1}}$	Observation
X_I	X_{I+1}	Y_I	Y_{I+1}		
1	2	1	4	3	$Y = X^2$ Is Not Linear Despite $R_{XY} = 0.97$
10	11	100	121	21	
		Y= $^1/_X$			
1	2	1	0.5	0.5	$Y = {}^1/_X$ Is Not Linear Despite $R_{XY} = -0.65$
10	11	0.1	0.090909	0.01389	
		Y = Log_{10}^{X}			Y = Log_{10}^{X} Is Not Linear Despite $R_{XY} = 0.92$
1	2	0	0.30103	0.30103	
10	11	1.0	1.04139	0.04139	
		$Y = \mathrm{Sinx}$			
1	2	0.017452	0.034899	0.01744709	$Y = \mathrm{Sinx}$ Is Almost Linear ($R_{XY} = 0.99$)
10	11	0.1736482	0.190809	0.01716082	
		$Y = \mathrm{Cosx}$			
1	2	0.99985	0.99939	0.00046	$Y = \mathrm{Cosx}$ Is Almost Linear ($R_{XY} = -0.97$)
10	11	0.98481	0.98163	0.00318	
$0 \le X \le 3.9$		$Y = \frac{1}{\sqrt{2\pi}} E^{\frac{-1}{2}X^2}$			
0.1	0.2	0.397	0.391	-0.06	$Y = \frac{1}{\sqrt{2\pi}} E^{\frac{-1}{2}X^2}$ Is Not Linear For $0 \le X \le 3.9$ ($R_{XY} = 0.93$)
1.1	1.2	0.2179	0.1942	-0.237	
$-3.9 \le X \le 3.9$		$Y = \frac{1}{\sqrt{2\pi}} E^{\frac{-1}{2}X^2}$			
-0.1	-0.2	0.397	0.391	0.06	$Y = \frac{1}{\sqrt{2\pi}} E^{\frac{-1}{2}X^2}$ Is Not Linear For $-3.9 \le X \le 3.9$ ($R_{XY} = 0.000361$)
1.1	1.2	0.2179	0.1942	-0.237	

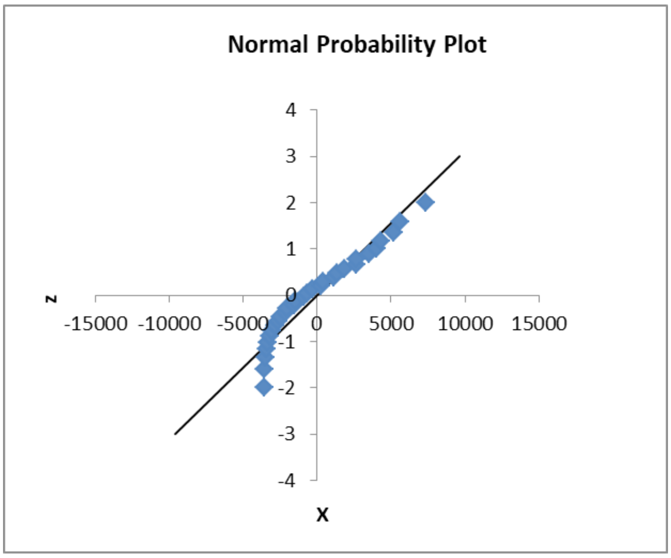

Figure 2 Normal probability plot of error score for X^3= α+βX.

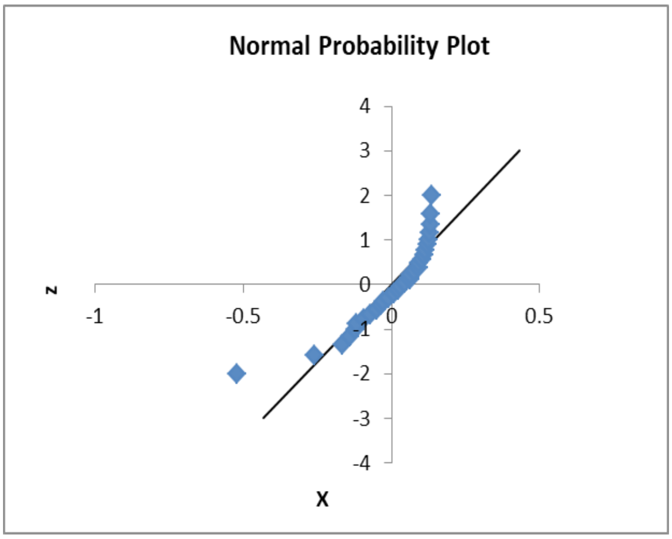

Figure 3 Normal probability plot of error score for〖log〗_10^X= α+βX.

of residuals and Y values. Y values are taken on the vertical Y-axis, and standardized residuals are plotted on the horizontal X- axis. A linear pattern of the scatter plot indicates that linearity assumption is met. There are other tests of normality like Shapiro-Wilk test, Kolmogorov-Smirnov test, etc. The Shapiro-Wilk test with greater power than the Kolmogorov-Smirnov test is preferred as a numerical means for assessing data normality [6]. Best is to undertake the Anderson – Darling test (AD-test) of normality which is an alternative to the chi-square and Kolmogorov-Smirnov (KS) goodness-of-fit tests. Power of AD-test is more than the same for Lilliefors test and KS test [7]. The AD-test can best be applied if there are no tied scores to test H_0: The univariate data follows normal distribution against H_1: The univariate data follows normal distribution. AD-test rejects the hypothesis of normality when the p-value is ≤ 0.05. Failing the normality test means that with 95% confidence the data does not fit the normal distribution. Passing the normality test implies no significant departure from normality. S_E and results of AD-test for normality of error scores of the chosen non-linear functions of X are shown in (**Table 3**).

Thus, the assumption of normality of error of prediction needs to be verified for fitting regression line and not merely the magnitude of correlation.

Take another example where X follows N (0, 1) and Y is the ordinate of N (0, 1) i.e.

Y=1/√2π e^((-1)/2 X^2). Clearly, X and Y are not linear.

Consider Case: 1 where 0≤X≤3.9 then r_XY= - 0.93302 and Case: 2 where -3.9≤X≤3.9 results in r_XY= 0.00036.

Interpretation of r_XY from Case: 1 is X and Y are highly correlated but correlation is negative i.e. increase of one unit in X will result in decrease of Y and vice versa. However, interpretation of r_XY from the Case – 2 will be just reverse. Low value of r_XY= 0.00036 tends to indicate that X and Y are independent, which is not the case in reality. In Case: 1, r_XY increased due to consideration of restricted range of values of X. In other words, truncated values of one or more variables (or homogeneity of data) may distort true relationships between two variables i.e. truncated score can underestimate or overestimate the correlation. However, many studies in social science involve variable (say X) taking positive values only assume the variable follows normal distribution and investigate relationship of X with other variables with homogenous sample and thus raise question about validity of such results [8].

The problem of truncated values may also occur if we want to find correlation between height and weight of students of say Class V. Here, r_XY will be poor, primarily due to range restriction of both the variables and also due to high homogeneity of the sample. Similarly, correlation between SAT scores and undergraduate grade point average (GPA) at some selective universities could be as low as 0.20 [8]. This is primarily due to small range of SAT scores of students admitted to the selective colleges and universities. Similarly, validity of selection test as a correlation between test scores and job performance is poor since range of test score is small for the persons selected through the test i. e. a homogeneous group. Other factors being equal, a restricted range usually yields a smaller correlation since it fails to reflect all the characteristics of the variable(s) being analyzed. Heteroscedasticity may be a serious empirical problem in truncated-sample models [9-11].

Even if linearity between variables is established, question arises regarding choice of independent and dependent variable when there is no cause and effect relationship. For example, Export Performance (EP) and GDP are having strong correlation. But, regression line of EP on GDP is different from regression of GDP on EP. Usually, choice of regression line is made depending on the purpose. To investigate equivalency between 5-point and 7-point scales, Colman [12], used linear regression equations, X_7= α_1+β_1 X_5 and X_5= α_2+β_2 X_7. But S_E of the two regressions are different. Probable solution could be to transform the variables so that variances of the transformed variables are same. One simple way to achieve this is to transform the original variables X and Y to P=X/(SD(X)) and Q=Y/(SD(Y)) .

This result in Var (P) = Var (Q) =1 i.e. homoscedasticity Regression coefficient for P on Q = same for Q on P and is equal to r_PQ=r_XY. In other words, slope of the two regression lines P on Q and Q on P coincide and the two regression lines are parallel [13].

For example, let X and Y are the scores on a 5-point scale and 7-point scale respectively of 100 individuals who responded to both the scales. Let us transform X to P where P=X/(SD(X)). Similarly, Q is obtained from Y by Q = Y/(SD(Y)). Details are shown in (**Table 4**).

Clearly, regression equation P=a+bQ is parallel to the regression equation Q=c+dP

where b=d, Efficiency or goodness of fit of both the regression equations are same since standard error of prediction of P from Q is equal to the same for Q from P = 0.988038. Thus, the transformations allow us to consider any of the variables P and Q as independent variable [14].

For the purpose of prediction, estimated values of P and Q can be transformed back to corresponding values of X and Y.

For multiple linear regression, such transformations to the dependent variable(Y) and each independent variable (X_i) will result in situation where each β_i = r_(X_i Y)

Other transforms

Equality of mean and variance

Consider a bivariate data on two variables X and Y with sample size n

Table 3. Anderson Darling test of Normality of error scores.

Predicting f(X)With X As The Independent Variable	SD Of Error Score (S_E)	Anderson – Darling Test		
		AD-Statistic	P-Value	Remarks For Error Score
F(X)=X^2	68.0392	1.048	0.007956	Normality Was Rejected
F(X)=X^3	3205.454	0.916	0.019875	Normality Was Rejected
$F(X) = Log_{10}^{X}$	0.14419	1.336	0.001834	Normality Was Rejected
F(X)=Cos X	0.009938	1.055	0.00904	Normality Was Rejected
F(X)=Sin X	0.002791	0.917	0.01975	Normality Was Rejected

Table 4. Regression with transformed variables.

	X	Y	$P = \frac{X}{SD(X)}$	Q= $\frac{Y}{SD(Y)}$
Mean	19.16	25.38	7.33533	6.983584
Variance	6.822626	13.20768	1	1
SD	2.612016	3.634237	1	1
Correlation	0.154207		0.154207	
Slope (Beta) For Regression Lines	0.214556 (Y On X) 0.110832 (X On Y)		0.154207 (Q On P) 0.154207 (P On Q)	
Intercept (Alpha) For Regression Lines	21.2691 (Y On X) 16.34707 (X On Y)		5.852425 (Q On P) 6.258413 (P On Q)	
Standard Error Of Prediction (S_E)	3.590766 (Y On X) 2.580772 (X On Y)		0.988038 (Q On P) 0.988038 (P On Q)	

Table 5. Correlation between X and Trigonometric functions of X.

Description of Variable(S)	$r_{X,Y}$	$r_{X,SinX}$	$r_{X,CosX}$	$r_{SinX,CosX}$
X ~N(0, 1) And $-3.9 \le X \le 3.9$	0.000361	1.0	0.001036	0.001037
X ~N(0, 1) And $0 \le X \le 3.9$	-0.933018	1.0	-0.96674	-0.96666
X: Integer Values Between 1 And 30		0.99982	-0.97156	-0.96696
X: Score In 5-Point Scale And Y: Score In 7-Point Scale	0.154207	0.999951	-0.99649	-0.99562

Define W_i= X_i- X ̅+ S_X   and P_i= Y_i- Y ̅+ S_Y

So ∑〖W_i 〗= nS_X   ⟹ W ̅= S_X

Now W_i- W ̅= X_i- X ̅ ⟹Var(W)= Var(X)

Thus, SD (W)= SD(X)= W ̅

Thus, mean (W) = variance (W) and ||W||^2= 2nS_X^2   ⟹ ||W|| =√2n S_X

Similarly, Mean (P)= SD(P) and ||P|| =√2n S_Y and r_PW=r_XY

β for W on P = r_PW.SD(P)/SD(W) = rPW.P ̅/(W ̅)

Thus, standard error for W on P is S_W √(1-r^2) will be lesser than standard error of P on W if W ̅<P ̅

Transformations to increase correlation

If r_XY is poor, it is possible to use transformation f(X) and/or g(Y) so that r_(f(X),g(Y)) is improved. Geometrically, it amounts to attempt to make the scatter plot of f(X) and g(Y) rather linear i.e. to achieve linearity. f(X) and g(Y) may be so chosen to ensure similarity of form of distribution of f(X) and g(Y). Clearly, f(X) and g(Y) will be non-linear [15].

Illustrative examples of transformations on a single variable are

1. Logarithmic functions: f(X)=log⁡X where X≥0,

Logarithms are inverses of exponential functions and can even change direction of correlation. It helps to reduce skewness.

2. Square root functions: f(X)= √X where X≥0

It has a moderate effect in change of shape of the distribution. It helps reducing right skewness.   3. Reciprocal function: f(X) = 1/X where X≠0

It changes shape of distribution and reverses order among values with same sign.

4. Trigonometric functions: f(X)=Sin X or g(X)=Cos X

Correlation between X and f(X) and g(X) for in illustrative cases are shown in (**Table 5**)

Almost perfect correlation was observed for f(X)=Sin⁡X even when r_XY ≈0 and also when the variable takes negative and positive values.|r_(X,CosX) |≈1. However, such empirical findings need to be rooted with theoretical explanations establishing high correlation between X and trigonometric function of X.

5. Arcsine Transformation: f(X)= 〖Sin〗^(-1) √X where 0≤X≤1 and f(X) is in radians range from -π/2 to π/2. It essentially stretches the tails of data. This is commonly used for proportions, like proportion of individuals in different genders [16].

6. Box-Cox transformation: f(X)= (X^λ-1)/λ   if λ≠0

= log⁡X if λ=0

In the Box-Cox linearity plot, r_(f(X),Y) is taken along the Y-axis for a given value of λ and λ are represented along the X-axis. The optimal value of λ is the one which corresponds to the maximum correlation (or minimum for negative correlation) on the plot. Wessa, (2012) has given software for Box-Cox plot. The transformation is used to reduce extent of non-normality and heteroscedasticity.

Attempts can be made to find class of functions so that r_XY = 〖Cosθ〗_xy=(x^T y)/||x||||y|| = 1 where x and y are deviation scores defined as x_i=X_i-X ̅ and y_i=Y_i-Y ̅. Assume θ_xy≠0 so that r_XY≠1

The condition requires x^T y= ||x||||y||

⟹x.[x^T y]= ||x||||y||. x ⟹ x.x^T [y]= ||x||||y||. x ⟹ A.[y]= ||x||||y||. x (1)

Where the matrix A= x.x^T

Note that A is a square matrix of order n×n with rank 1. Thus, A^(-1) does not exist. However, one can find generalized inverse (G-inverse) of the matrix A. Let it be denoted by G_(n×n)where AGA = A

From the above equation, it follows that y=G.||x||||y||. x   (2)

⟹y/||y|| = G.||x||. x

Estimated value of the Y-vector (Y ̂) will be perfectly correlated with X. Test of linearity will follow.

Conclusion

Since G-inverse is not unique, solution of (2) is not unique. Moore-Penrose method of finding G-inverse is popular. Solution of the equation (2) will give a method to introduce linearity between two non-linear variables and can help to convert non-linear relations to linear relationships. Such solution may be extended to ensure linearity between a dependent variable (Y) and a set of independent variables (Multiple linear regressions) or between set of dependent variables and set of independent variables (Cannonical regression). Empirical illustration of G-inverse and extensions for multiple linear regressions and Cannonical regressions are suggested for future studies.

References

1 Abrami PC, Cholmsky P, Gordon R (2001) Statistical analysis for the social sciences: An Interactive Approach. Needham Heights MA: Allyn Bacon.

2 Bobko P (1995) Correlation and Regression: Principles and Applications for Industrial Organizational Psychology and Management, New York: Mc.Graw-Hill.

3 Chen PY, Popovich PM (2002) Correlation Parametric and Nonparametric Measures. Sage University Papers Series on Quantitative Applications in the Social Sciences. Thousand Oaks CA: Sage 07-139.

4 Colman AM, Norris CE, Preston CC (1997) Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales Psychological Reports 80: 355-362.

5 Das KR, Rahmatullah Imon AHM (2016) A brief review of tests for normality. Am J Theor Appl Stat 5: 5-12.

6 Eisenbach R, Falk R (1984) Association between Two Variables Measured as Proportion of Loss Reduction. Teaching Statistics 6:47-52.

7 Falk R, Well AD (1997) Many Faces of the Correlation Coefficient. J Stat Edu 5: 1-12.

8 Field AP (2003) Can meta-analysis be trusted? Psychologist 16: 642-645.

9 Goodwin LD, Leech NL (2006) Understanding Correlation: Factors that Affect the Size. J Exp Edu 74: 251-266.

10 Loco JV, Elskens M, Croux C, Beernaert H (2002) Linearity of calibration curves: use and misuse of the correlation coefficient. Accred Quality Assur 7:281-285.

11 Pedhazur EJ (1973) Multiple Regression in Behavioural Research. New York: Holt, Rinehart & Winston.

12 Rodgers JL, Nicewander WL (1988) Thirteen ways to look at the correlation coefficient. The Amer Stat 42: 59-66.

13 Rovine MJ, von Eye a (1997) 14th way to look at a correlation coefficient: Correlation as the proportion of matches. Amer Stat 51:42-46.

14 Vaughan ED (1998) Statistics Tools for understanding data in the behavioural sciences. Upper Saddle River NJ: Prentice-Hall.

15 Vetter TR (2017) Fundamentals of research data and variables: the devil is in the details. Anesth Analg 125: 1375-1380.

16 Wessa P (2012) Box-Cox Linearity Plot (v1.0.5) in Free Statistics Software (v1.1.23-r7), Office for Research Development and Education.