İstatistik Araştırma Dergisi
Journal of Statistical Research

ISSN: 2791-7614

(2025), 15 (2)
Başvuru / Received :12.03.2025
Kabul / Accepted :21.10.2025

https://dergipark.org.tr/tr/pub/jsstr

ARAŞTIRMA MAKALESİ **RESEARCH ARTICLE**

Transformation to Achieve Perfect Correlation

Satyendra Nath CHAKRABARTTY
Indian Ports Association, Indian Statistical Institute, Indian Maritime University
chakrabarttysatyendra3139@gmail.com
Orcid No: 0000-0002-7687-5044

Anish CHAKRABARTY
Indian Statistical Institute
chakrabarty.anish@gmail.com
Orcid No: 0000-0002-1993-2006

Abstract

Correlation and linear regression are common means to evaluate association and empirical relationships between two or more variables. Such relationships often show significant departure of $|r_{XY}|$ from unity. Existing transformations to increase correlation fail to achieve perfect correlation. For a bivariate data, the paper proposes transforming Y to $y = G.\|x\|\|y\|$, which gives $r_{X\,y} = 1$ where G is the G-inverse of the matrix $A = x.x^T$ and x, y denote vectors of deviation scores. The concept is extended to perfect 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 (Canonical regression), avoiding problems of insignificant beta coefficients in univariate and multivariate regression models and outliers. Empirical illustration of G-inverse and extensions for multiple linear regressions and Canonical regressions are also given. The proposed transformation is a novel method of introducing perfect correlation between two variables. Extension of the concept in multiple linear regressions and canonical regression will go a long way in empirical researches in various branches of science. Future studies may include finding distribution of the proposed perfect correlations and comparison of efficacy of our suggested approach against other traditional ones by providing quantitative evidences.

Keywords: Correlation coefficient, Generalized Inverse, Multiple Linear Regressions, Canonical Regression, Normal Distribution

Corresponding Author / Sorumlu Yazar: Satyendra Nath CHAKRABARTTY, Indian Ports Association, Indian Statistical Institute, Indian Maritime University.

Citation / Atıf: CHAKRABARTTY S.N., CHAKRABARTTY A. (2025). Transformation to Achieve Perfect Correlation. *İstatistik Araştırma Dergisi*, 15 (2), 1-12.

Citation / Atıf: CHAKRABARTTY S.N., CHAKRABARTTY A. (2025). Transformation to Achieve Perfect Correlation. *İstatistik Araştırma Dergisi*, 15 (2), 1-12.

Mükemmel Korelasyona Ulaşmak için Dönüşüm

Özet

Korelasyon ve doğrusal regresyon, iki veya daha fazla değişken arasındaki ilişkiyi ve ampirik ilişkileri değerlendirmek için yaygın araçlardır. Bu tür ilişkiler genellikle $|\boldsymbol{r_{XY}}|$'nin birlikten önemli ölçüde saptığını gösterir. Korelasyonu artırmak için yapılan mevcut dönüşümler mükemmel korelasyona ulaşmada başarısız olur. İki değişkenli veriler için, makale Y'yi $\boldsymbol{y} = \boldsymbol{G}.\|\boldsymbol{x}\|\|\boldsymbol{y}\|$'ye dönüştürmeyi önerir; bu da $\boldsymbol{r_{X\,y}} = \mathbf{1}$ 'i verir; burada G, $\boldsymbol{A} = \boldsymbol{x}.\boldsymbol{x^T}$ matrisinin G-tersidir ve x, y sapma puanlarının vektörlerini belirtir. Kavram, bağımlı değişken (Y) ile bağımsız değişkenler kümesi (Çoklu doğrusal regresyonlar) veya bağımlı değişkenler kümesi ile bağımsız değişkenler kümesi (Kanonik regresyon) arasındaki mükemmel doğrusallığa genişletilir ve tek değişkenli ve çok değişkenli regresyon modellerinde ve aykırı değerlerde önemsiz beta katsayıları sorunlarından kaçınılır. G-tersinin ampirik gösterimi ve çoklu doğrusal regresyonlar ve Kanonik regresyonlar için uzantılar da verilmiştir. Önerilen dönüşüm, iki değişken arasında mükemmel korelasyon tanıtmanın yeni bir yöntemidir. Kavramın çoklu doğrusal regresyonlarda ve kanonik regresyonda genişletilmesi, çeşitli bilim dallarındaki ampirik araştırmalarda uzun bir yol kat edecektir. Gelecekteki çalışmalar, önerilen mükemmel korelasyonların dağılımını bulmayı ve nicel kanıtlar sağlayarak önerilen yaklaşımımızın etkinliğinin diğer geleneksel yaklaşımlarla karşılaştırılmasını içerebilir.

Anahtar sözcükler: Korelasyon katsayısı, Genelleştirilmiş ters, Çoklu doğrusal regresyonlar, Kanonik regresyon, Normal dağılım

1. Introduction

Correlation, $r_{XY} \in [-1,1]$ taken as a measure of linear association between two variables, is vastly used in various fields of applied and theoretical research. Linear regression, reflecting the empirical relationship between a dependent variable (Y) and an independent one (X) works well in the presence of high value of correlation coefficient between X and Y i.e. $|r_{XY}|$. Given $n \in \mathbb{N}^+$ paired observations, the error scores $\epsilon_{YX} = y_i - (\alpha_1 + \beta_1 x_i)$ and $\epsilon_{XY} = x_i - (\alpha_2 + \beta_2 y_i)$ from the fitted regression equations are assumed to follow Gaussian distribution with zero mean, constant variance and uncorrelated with X and Y. However, regression equations of Y on X and that of X on Y are different, with different values of error variances. The quality or goodness of fit of a regression equation depends on the departure of $|r_{XY}|$ from unity. In real scenarios, relationships between variables rooted in domains such as economics, social sciences, medical sciences, etc. often turn out to be nonlinear; showing significant departure of $|r_{XY}|$ from unity. Loco et al. (2002) observed that, r_{XY} is not suitable for assessing linearity of calibration curves. If X and Y are nonlinearly related, the slope of the curve changes as does the value of one of the variables. In such a case, transformations are used on the variable X ($X' = f(X)$), or the variable Y ($Y' = g(Y)$), or both, to improve degree of linear association between the transformed variables i.e., $|r_{X'Y'}| > |r_{XY}|$ and still keeping the regression analysis appropriate. Several suggestions in the form of nonlinear transformations have been made over the years to increase the correlation between two variables. However, finding the best transformation given a pair of variables still needs a lot of prior diagnosis. Moreover, such transformations do not provide universal solution.

To fill this void, the paper proposes a simple data-based non-linear transformation using G-inverse that achieves exact linear alignment and the concept is extended to perfect 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 (Canonical regression). Empirical illustration of transformation involving G-inverse and extensions for multiple linear regressions and Canonical regressions are also given.

2. Related work

Finding suitable transformations that make a variable more linearly aligned with another has a fairly long history. Let us consider a brief exposition of the methods frequently used by statisticians. While linear transformations to a variable like $cX + b$, where$b, c \in \mathbb{R}$, preserve the extent of its linear relationship to another variable, non-linear functions were explored to improve value of the r_{XY}. Common class of such functions leads to power transforms, where variables are raised to a real power. Most common variants of the same include the *square root function*$f(X) = \sqrt{X}$, where $X \geq 0$. It has a moderate effect in changing the shape of distributions. However, it helps reducing right skewness. Another popular choice is *reciprocal function*,$f(X) = \frac{1}{X}$, for $X \neq 0$. It also changes the shape of the distribution and reverses order among values with same sign. *Logarithm function*,$f(X) = \log X$, $\forall X \geq 0$ are also popular choices and can even change the direction of correlation. It typically finds its usefulness mostly at reducing skewness. Kovacevic (2011) found that correlation between Life expectancy and Human Development Index (HDI) exceeded the same between Life expectancy and GDP but the inequality got reversed when logarithmic transformations were used. *Trigonometric functions* such as $f(X) = sin\, X$ or $g(X) = cos\, X$ tend to become useful at times in improving linear association. Chakrabartty (2023) presented examples where each of $|r_{X,sinY}|$, $|r_{X,cosY}|$ and $|r_{sinX,cosY}|$ was close to unity, even if *X* and *Y* share clear curved relationships. Moreover, high value of correlations did not support normality of error in prediction. Thus, the assumption of normality of residuals needs to be verified for fitting regression line along with verification of no correlations of explanatory variables with error scores, homoscedasticity and linearity between the response variable and the explanatory variable (Field & Wilcox, 2017; Yellowlees et al., 2016). However, in reality, all the four assumptions of linear regression are not satisfied (Erceg-Hurn & Mirosevich, 2008).

Given that X takes values between [0, 1], *arcsine transformation*:$f(X) = sin^{-1}\sqrt{X}$ becomes crucial, projecting the output in radians ranging between $-\frac{\pi}{2}$ to $\frac{\pi}{2}$. Perhaps, the most celebrated variance-stabilizing transforms is the *Box-Cox transformation* given by Box and Cox, (1964) as

$f(X) = \begin{cases} \frac{X^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log X & \text{if } \lambda = 0 \end{cases}$, (natural logarithm is taken for $\lambda = 0$). In the Box-Cox linearity plot, $r_{f(X),Y}$ is plotted in the Y-axis and values of the transformation parameter λ are taken along the X-axis. The value of λ which corresponds to the maximum correlation (minimumfor negative correlation) is taken as the optimal λ. We refer to Wessa (2012) for software support inthis regard, especially for Box-Cox plot.

It may be observed that, all the above methods avoid the contextual problem of dependency on to another (or a set of) variable(s). Consequently, there is no straight forward way of determining which transformation to employ given the set of variables. Also, no such transformations guarantee an exact linearalignment between the variables under consideration, post-translation. In contrast, this paper proposes a class of functions to achieve perfect correlation i.e.

$r_{XY} = cos\ \theta_{xy} = \frac{x^T y}{\|x\|\|y\|} = 1$,

using generalized inverse (G-inverse) of the matrix $\mathbf{A} = x.x^T$, where deviation scores *x* and *y* are defined as $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$.

Assuming $\theta_{xy} \neq 0$ implying$r_{XY} \neq 1$, the above said condition requires

$$x^T y = \|x\|\|y\|$$

$\Longrightarrow x.[\, x^T y] = \|x\|\|y\|.\, x$

$\Longrightarrow\ x.x^T[y] = \|x\|\|y\|.\, x$

$\Longrightarrow\ A.[y] = \|x\|\|y\|.\, x$ (1)

where the matrix $\mathbf{A} = x.x^T$ is of order $n \times n$ has rank 1. Thus, $\boldsymbol{A^{-1}}$ does not exist. However, one can find G-inverse of the matrix **A** as $\boldsymbol{G}_{n\times n}$ where **AGA = A** and **GAG = G**

From the above equation, it follows that

$y = \boldsymbol{G}.\|x\|\|y\|.\, x$

$$\mathcal{y} \Longrightarrow \frac{y}{\|y\|} = \boldsymbol{G}.\|x\|.\, x. \quad (2)$$

Transformed value of the Y-vector ($\mathcal{y}$) will be perfectly correlated with X. Test of linearity can also be undertaken.

It may be noted that solution of (2) is not unique since G-inverse of a given matrix 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.

***Example 1*:** Our first example demonstrating the efficacy of the proposed method comes as a simulated experiment. We randomly select n = 500 samples from $N(0, 1)$ under the constraint

$|X| \leq 3.9$. Let us define the corresponding ordinates as Y. Also, let us denote the deviation scores as $x = X$ - Mean(X). Similarly, denote $y = Y - Mean\ (Y)$. Descriptive statistics corresponding to the dataset are shown in table 1.

Table 1. Descriptive statistics.

Description	**Original variables**		**Deviation scores**		$\psi = \boldsymbol{G}.\|x\|\|y\|.x$
	X	Y	x	y	
Mean	-0.0586497	0.2784992	0	0	≈0
Variance	1.069737	0.0132699	1.069737	0.0132699	0.0132699
CV = $\frac{SD}{Mean}$	-17.63489	0.4136279			
Correlation	r_{XY} = 0.05188477		r_{xy} = 0.05188477		$r_{x\psi}$ = 1

Clearly, $\psi = \boldsymbol{G}.\|x\|\|y\|.x$ resulted in $r_{X\psi} = 1$, even if $r_{XY} = 0.05188$. The residuals for the regression Y = α+ $\beta[\boldsymbol{G}.\|x\|\|y\|.x]$ and the scatter plot of X and ψ are depicted in Figure 1 and 2 respectively.

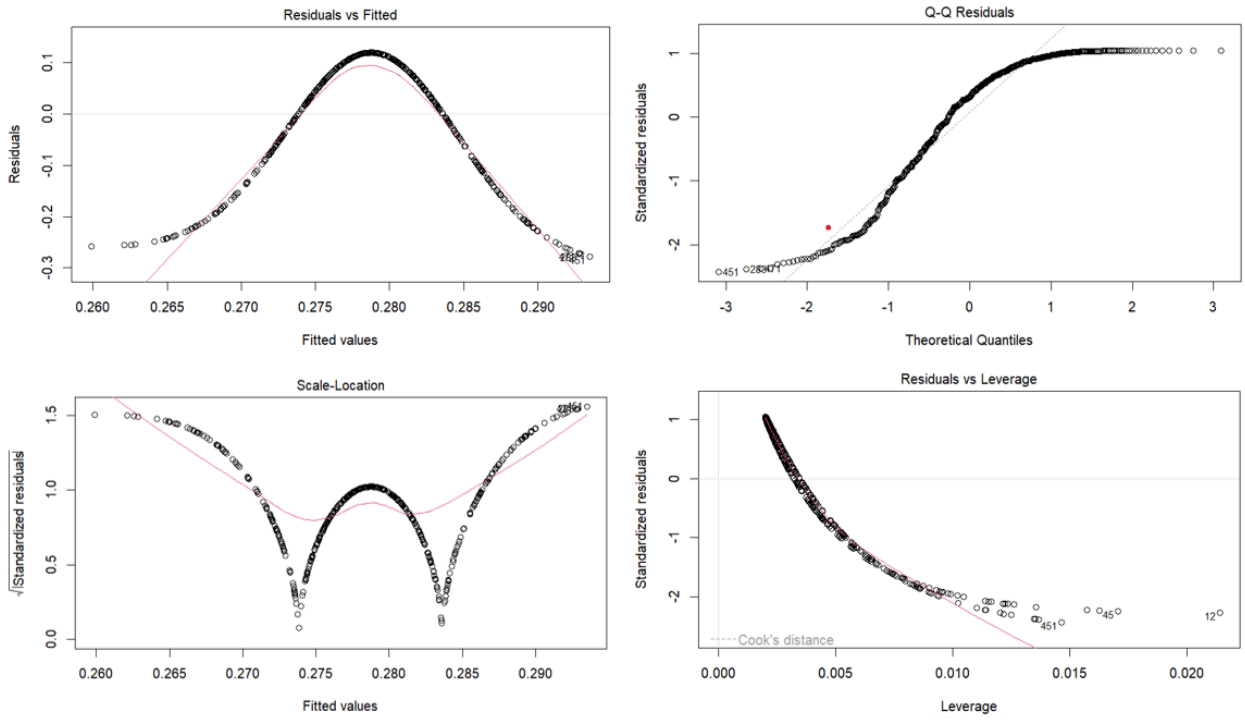

Figure 1. Scatter plots and corresponding residuals of Y=α + $\beta[\boldsymbol{G}.\|x\|\|y\|.x$ (Correlation = 1) The plot of X and $\psi = G.\|x\|\|y\|.x$ ($r_{X\psi}$=1) reveals the effectiveness of the transformation. It showcases the perfect linear relation between the two variables.

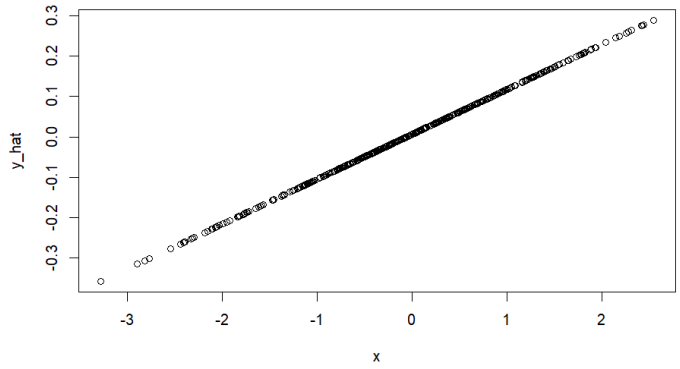

Figure 2. Plot of X and $\mathcal{y} = G.\|x\|\|y\|.x$

Here, regression equation of $\mathcal{y}$ on x is given by $\hat{\mathcal{y}} = 1.114e\text{-}01x$ where SD of the residuals ≈ 0. Similarly, $\hat{x}$ on $\mathcal{y}$ is obtained as $\hat{x} = 8.979\ \mathcal{y}$

However, presence of outliers can affect correlation and mislead the nature of the association among the variables considered (Kim et al., 2015; Niven & Deutsch, 2012). Outliers are data points which deviate from the majority of the data points and weaken the correlation making the data more scattered. Outliers in data may indicate violation of the assumptions of normality and homogeneity and thus, impact the result of linear regression like inaccurate intervals, lowering of statistical power, Type 1 and Type 2 errors (Brossart et al., 2011). In addition, departure from bivariate normality may distort associations as illustrated by Wilcox (2022).

Among the bivariate points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, if X_i is an outlier, the point (X_i, Y_i) is a leverage point. For the linear regression equation fitted with the data points be $Y = \alpha + \beta X$ where

$r_i = Y_i - \alpha - \beta X_i$ denotes the residual for the *i*-th observation. If r_i is an outlier among the set of residuals $\{r_1, r_2, \ldots \ldots, r_n\}$ and the corresponding (X_i, Y_i) is a leverage point, then the point (X_i, Y_i) is a bad leverage point which can negatively impact correlation coefficient and can result in a poor fit of the linear model. If (X_i, Y_i) is a leverage point but the corresponding r_i is not an outlier, the point (X_i, Y_i) is a good leverage point and is taken as consistent with the regression equation$Y = \alpha + \beta X$.

Methods have been proposed for detecting the bad leverage points of regression line, using the least median of squares (LMS) estimator (Rousseeuw and van Zomeren, 1990), Theil–Sen estimator compute the slope (β) as median of the$S_{ij}{}'s$ where $S_{ij} = \frac{Y_i - Y_j}{X_i - X_j}$ for each $i < j$ and the intercept $\alpha = M_Y - \beta M_X$, where M_Y is the median of $\{Y_1, Y_2, \ldots \ldots, Y_n\}$ and M_X is the median of $\{X_1, X_2, \ldots \ldots, X_n\}$. Wilcox (2023) proposed computation of slope and intercept removing the bad leverage points, which result in loss of information.

Regression equation of the form$y = f(x) + e$, where $f(x)$ is the regression function, e is error, and $r_{f(x),e} = 0$ give $r_{y,e} = \frac{SD(e)}{SD(y)} = \sqrt{1 - r_{xy}{}^2}$. Thus, perfect correlation $(2) \Longrightarrow \frac{SD(e)}{Sd(y)} = 0 \Longrightarrow$SD (e) =0 i.e. predicting 100% of the variance of y by the regression line with no residual variance, i.e. perfect correlation can be taken as zero impact of outliers. In the example at Table- 1, an outlier *x*: Q3 + (1.5 * IQR) = 2.707498 and Q1 – (1.5 * IQR) = -2.670538 where IQR = Q3 – Q1. Predicted value of $\widehat{\mathcal{y}_i}$

for x_i >2.707498 also resulted in an outlier. The same goes for lower limit. This is expected since $\hat{\mathcal{y}}$ and *x* exhibit an exact linear relationship and $\mathcal{y} = \alpha + \beta x$ produces SD (error) ≈ 0.

3. Multiple correlations

Multiple correlation coefficient depicts the maximum linear relationship (correlation) between a variable and a linear function of a set of independent variables. It is intrinsically linked to the multiple linear regression (MLR) setups. Here, a dependent variable Y is predicted based on a set of independent variables. In our experiment, we use the 'Airfoil Self-Noise Dataset', consisting of 1503 observations on six real valued attributes (Brooks et al. 2014). A brief data description goes as follows: Frequency, in Hertz (X_1), Angle of attack, in degrees (X_2), Chord length, in meters (X_3), Free-stream velocity, in meters per second (X_4), Suction side displacement thickness, in meters (X_5) and the response variable, scaled sound pressure level, in decibels (Y). Let the vector $C =$

$(r_{x_1y}, r_{x_2y}, \dots r_{x_ny})^T$ consists of the correlations between original variables of the response and the explanatory variables. Using the unaltered observations, we obtain:

$C = (-0.3907114, -0.1561075, -0.2361615,\ 0.1251028, -0.3126695)^T$

The correlation matrix of order 5 ×5 corresponding to the independent variables $X_1, X_2, X_3, X_4\ and\ X_5$ was as follows.

$$R_{XX} = \begin{pmatrix} 1 & -0.27276 & -0.00366 & 0.133664 & -0.23011 \\ -0.27276 & 1 & -0.50487 & 0.05876 & 0.753394 \\ -0.00366 & -0.50487 & 1 & 0.003787 & -0.22084 \\ 0.133664 & 0.05876 & 0.003787 & 1 & -0.00397 \\ -0.23011 & 0.753394 & -0.22084 & -0.00397 & 1 \end{pmatrix}$$

As such, the multiple correlation coefficient turns out to be $R^2 = C^T {R_{XX}}^{-1} C = 0.5157097$. In addition, we validated this result based on the MLR of Y based on $(X_1, \dots, X_5)$, which also comes out to be 0.5157. Clearly, improving and implementing our proposed transformation (as given in equation 2) on each explanatory variable X_i separately will not be adequate. This is based on the observation that if all r_{x_iy}'s are made equal to 1, the multiple correlation comes out to be the sum of the entries of the inverse of the following correlation matrix

$${R_{XX}}^{-1} = \begin{pmatrix} 1.144444 & 0.472385 & 0.234194 & -0.18178 & -0.04155 \\ 0.472385 & 3.441658 & 1.252592 & -0.27889 & -2.20871 \\ 0.234194 & 1.252592 & 1.510754 & -0.11284 & -0.55662 \\ -0.18178 & -0.27889 & -0.11284 & 1.041698 & 0.147507 \\ -0.04155 & -2.20871 & -0.55662 & 0.147507 & 2.532127 \end{pmatrix}$$

which is greater than 1. In the absence of scaling, this violates the basic property of R^2 being the fraction of variance of the dependent variable that is explained by the independent ones. Instead,observe that for the given explanatory variables $X_1, X_2, \dots, X_m$ and response Y, similar to the earlier case, we require

$x_i.{x_i}^T[y] = \|x_i\|\|y\|.\ x_i\ \ \forall i = 1,2, \dots, m.$

As such,

$$(\sum_{i=1}^{m} x_i.{x_i}^T)[y] = \sum_{i=1}^{m} \|x_i\|\|y\|.x_i \implies \mathcal{y} = G^m \sum_{i=1}^{m} \|x_i\|\|y\|.x_i \qquad (3)$$

where $G^m = \{x_1.{x_1}^T + x_2.{x_2}^T + \cdots + x_m.{x_m}^T\}^-$ ($^-$ being the notation for G-inverse).

Let us replace the vector $\boldsymbol{C} = (r_{x_1y}, r_{x_2y}, \dots r_{x_ny})^T$ for raw data by $\boldsymbol{C'} = (r_{x_1\mathcal{y}}, r_{x_2\mathcal{y}}, \dots, r_{x_m\mathcal{y}})^T$. Based on the dataset at hand, we obtain

$$C' = (0.3748757, 0.3748756, 0.3747865, 0.3748757, 0.3951889)^T$$

That in turn gives ${C'}^T {R_{XX}}^{-1} C' = 1$. We also validated our findings by regressing $\mathcal{y}$ on $X_1, X_2, \dots, X_5$ using an additive linear regression model (say, L^*) . Considering the transformation, it is quite straight forward to conclude that the fitted model passes all tests of efficacy at 5% level of significance. We also checked for the normality of the residuals obtained from the fitted model. The Anderson-Darling test of normality at the same level of significance results in a positive outcome, i.e. the residuals are indeed standard Gaussian. The figure 3 given below gives a visual representation of the randomness in the residuals obtained, having zero mean.

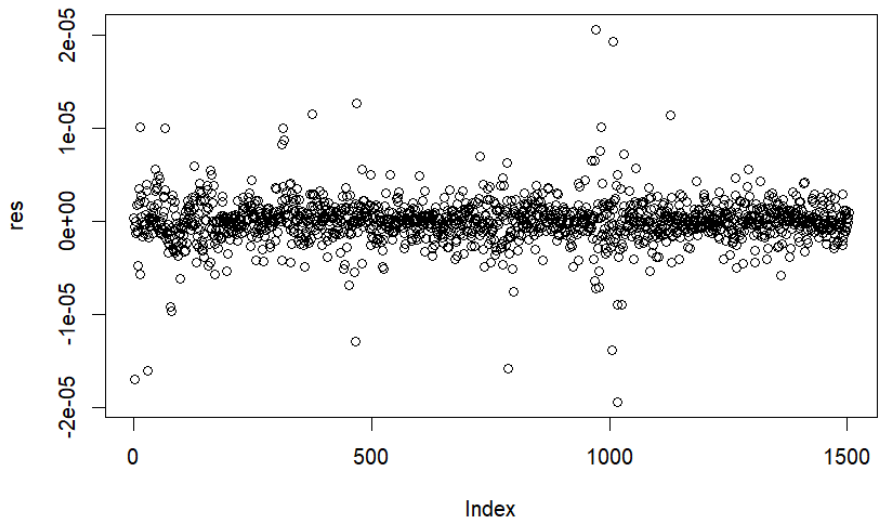

Figure 3. Plot of residuals from the model $L*$

In addition to presence of outliers, multiple linear regression may give rise to paradoxical findings. For example, as per the univariate regression of Y on the predictorX_1, the significant regression coefficient β_{YX_1}may turn out to be insignificant upon consideration of additional independent variables; similarly, a significant β-coefficient in multiple regression may not be significant in the univariate regression (Feng et al. 2016). These are primarily due to different values of r_{Y,X_i} and non-satisfaction of the assumptions of the univariate and multiple regression models. Similar problems were indicated by Agresti, (2002) in logistic regression for binary outcome, log-linear regression for counting data and Cox proportional hazards regression for survival analysis (Cox, 1972).

For perfect multiple correlation, $R^2 = 1 \Longrightarrow \frac{\Sigma_{i=1}^{n} e_i^2}{\Sigma_{i=1}^{n}(y_i - \hat{y})^2} = 0$, implies no residual variance, i.e. zero impact of outliers. The property along with r_{y,X_i}= 1 avoid the problems of insignificant beta coefficients in univariate and multivariate regression models and outliers.

For the data considered on 'Airfoil Self-Noise Dataset' with 1503 observations, the obtained multiple regression equation was

$\hat{y} = -4.612e^{-8} + 0.003577x_1 + 2.983x_2 + 169.4x_3 + 0.2763x_4 + 5.3x_5$

with SD of residuals = 0.000002421 ≈ 0

Outlier limits for each x_i and $\hat{y}$ obtained from the above equation are given in Table 2.

Table 2. Outliers in Multiple Linear Equation.

	x_1	x_2	x_3	x_4	x_5	$\hat{y}$
Q3 + (1.5 * IQR)	5913.619	14.9677	0.3587518	67.98925	0.0239972	48.12163
Q1 – (1.5 * IQR)	-6886.381	-16.6323	-0.352448	-58.81075	-0.0281659	-48.34267

As such, if observations corresponding to all explanatory variables are simultaneously outlying following the same trend (above or below the Tukey thresholds), predicted values would also be so. This property may not be satisfied by other combinations of the variables.

4. Canonical Correlation

Another frontier where the proposed transformation can be useful is the canonical correlation analysis (CCA). It deals with the linear association between two sets of variables, say a list of labtest results and a set of clinical observations on patients. To put it mathematically, consider data on n- observations on a collection of vectors $X = (X_1, X_2, \dots, X_p)^T$ having p-attributes and $Y = (Y_1, Y_2, \dots, Y_q)^T$ for $p, q > 1$. The underlying objective is to find real vectors $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^q$ such that the correlation between $\sum_{i=1}^{p} a_i X_i$ and $\sum_{i=1}^{q} b_i Y_i$ is maximized. The optimal inner products thus obtained are called the primary canonical variates. We can find up to min $\{p, q\}$ variates, constraining the latter ones to be uncorrelated to the previously found. Ideally, this is similar in spirit to MLR, except that number of attributes ($q \geq 1$) corresponding to *Y* is increased. As such, our method should follow suit. For a detailed description of CCA one may turn to Mardia et al. (1982). Song-Gui and Shein-Chung (1987) have shown that under a simple condition, measures of multivariate association in terms of canonical correlations is equal to one if and only if there exists a linear relationship between the sets of variables.

Vasylieva et al. (2023) observed less uses of canonical analysis in Medicine (13.85%) and Social Sciences (5.32%). Canonical correlation analysis were used for testing specificity of environmental risk factors for developmental outcomes (Bignardi et al. 2022); relationships between physical strength measurements and anthropometric variables (Malakar et al. 2022); health status and economic growth in terms of gender-oriented inequalities (Gavurova et al. 2020). Through canonical correlation analysis, Fox and Hammond, (2017) found three significant canonical functions between multidimensional Barratt Impulsiveness Scale (BIS-11) and multidimensional Psychopathic Personality Inventory (PPI) with about 70% of the variance shared between the sub-scales. Explanatory power of the multi-dimensional relationship between impulsivity and psychopathy was much higher than the simple impulsivity-psychopathy correlation. However, interpretation of different canonical functions giving three different patterns of relationship between the PPI and BIS 285 needs caution. To evaluate the relationships between quality of life (QOL) of patients suffering from chronic obstructive pulmonary disease (COPD), Liu et al. (2022) computed canonical correlation between four measures of QOL and 15- clinical objective indicators and found two pairs of canonical variables accounting for 45.8% and 33.8% of the variance, respectively were statistically significant. However, predictions of dependent variables using canonical correlations are rather rare.

To demonstrate our proposition, let us consider the dataset 'Avila' that has been extracted from 800 images of the 'Avila Bible', a giant Latin copy of the whole Bible produced during the XII century between Italy and Spain (Stefano et al. 2018). The two underlying classes of attributes we have observations on are namely:

Inter-columnar distance (X_1), upper margin (X_2), lower margin (X_3), exploitation (Y_1), row number (Y_2), modular ratio (Y_3), interlinear spacing (Y_4), and weight (Y_5).

A quick implementation of the CCA on this dataset using *R* reveals the maximum correlation to be CC_1= 0.8715930, with corresponding factor loadings (F_X and F_Y) given as

$$CC = (0.8715930, 0.3314671, 0.2307935)^T,$$

$$F_X = \begin{pmatrix} 0.0259 & 0.0875 & -0.2118 \\ -0.0324 & -0.0151 & -0.0447 \\ -0.0043 & -0.0367 & -0.0115 \end{pmatrix}, \text{ and}$$

$$F_Y = \begin{pmatrix} -0.0044 & -0.0609 & 0.0088 & -0.0190 & -0.0081 \\ 0.0632 & 0.0335 & 0.0788 & -0.0340 & 0.0402 \\ -0.0006 & -0.0049 & 0.0242 & 0.0361 & -0.0029 \\ 0.0131 & -0.0130 & -0.0469 & 0.0164 & 0.0009 \\ -0.0096 & 0.0104 & 0.0172 & 0.0188 & 0.0619 \end{pmatrix}$$

To propose an improvement plan, let us first look at the pairwise correlation between the components of X and Y variables.

$$r_X = \begin{pmatrix} 1 & -0.7626 & 0.2027 \\ -0.7626 & 1 & -0.1836 \\ 0.2027 & -0.1836 & 1 \end{pmatrix},$$

$$r_Y = \begin{pmatrix} 1 & 0.1278 & -0.0051 & 0.1205 & 0.2396 \\ 0.1278 & 1 & 0.1698 & -0.6092 & -0.3028 \\ -0.0051 & 0.1698 & 1 & 0.2300 & -0.2932 \\ 0.1205 & 0.6092 & 0.2300 & 1 & -0.0699 \\ 0.2396 & -0.3028 & -0.2932 & -0.0699 & 1 \end{pmatrix}, \text{ and}$$

$$r_{X.Y} = \begin{pmatrix} -0.0344 & 0.6532 & 0.1074 & 0.5996 & -0.2958 \\ -0.0637 & -0.8281 & -0.2114 & -0.6608 & 0.3543 \\ 0.2723 & 0.0522 & 0.0364 & 0.1703 & 0.0102 \end{pmatrix}$$

It may be observed that in the instant case, the Y-variables are not so strongly correlated amongst themselves. We emphasize that our formula for improved multiple correlations (3) is not really needed since aligning any one of $Y_j's$ with arbitrary X_i would yield canonical correlation = 1. However, our goal is to maximize the subsequent CC values as well. This poses the challenge regarding a suitable choice of the pair. Our prescription is to first choose the Y_{j^*} such that

$$j^* = \underset{j=1,2,\ldots,q}{\operatorname{argmin}} \sum_{i=1}^{p} \left| r_{x_i y_j} \right|$$

In other words, the variable Y_j that has the least cumulative linear association to X variables. If we align the same more towards the X point cloud, the overall association would tend to increase. Now, apply the transformation (2) on Y_{j^*} based on the X_{i^*} which has the least absolute correlation, i.e. $\min_{i=1,2,\ldots,p} \left| r_{x_i y_{j^*}} \right|$. As such, similar to that in case of MLR, we establish a perfect linear canonical correlation between the two set of variates. Now, execute the usual CCA with unaltered X and the newly transformed Y. Also, as we transform only the Y_j which has the least overall effect on$X$'s, the second CC is least sacrificed in the process.

Based on the dataset at hand, the selected pair of variables turn out to be (Y_3, X_3) (see$r_{X,Y}$; also due to the values of $\sum_{i=1}^{3} \left| r_{x_i y_j} \right|$, $j = 1, \ldots, 5$ as

$(0.3704, 1.5336, 0.3552, 1.4307, 0.6603)$.

Calculate $\widehat{Y_3}$ using the transformation (2). Now, running the CCA on R based on the altered data we obtain:

$CC' = (1, 0.8708266, 0.2102745)^T$

As such, the two sets of variables are perfectly aligned post-transformation, with minuscule degradation to subsequent canonical correlations.

5. Discussion

The proposed transformation gives a computationally efficient way to establish perfect linear association. Due to abundant guidance available in the form of packages to compute G-inverse, the method becomes easy to implement. In addition, the inherent generalizability of the method for multiple variables makes it suitable for multiple regressions, in general and particularly canonical correlation, a sphere which hardly has seen any such modification suggested to it.

6. Limitations

However, the improvements come with some limitations. An immediate question which may arise is regarding the choice of the G-inverse of the given matrix. In fact, for a matrix **A**= $x.x^T$, the entire class of generalized inverse can be generated from a given G-inverse **G** by $\boldsymbol{G + U - GAUAG}$ where **U** is arbitrary or by $\boldsymbol{G + V(I - AG) + (I - GA)W}$ for arbitrary matrix $\boldsymbol{V}$ and $\boldsymbol{W}$ (Rao and Mitra, 1971).

In our experiments, we use in particular the Moore-Penrose inverse primarily due to its computational ease. While other choices are also viable, the non-uniqueness of it imposes the same property on to the transformed observations. In this regard, we prescribe selecting the transformation resulting in the least variance. A detailed study aiming to investigate the most efficient inversion technique might be taken up as a future endeavor.

Other limitations include non-consideration of population heterogeneity. It may not be realistic to assume that a single-population model can explain all kinds of individual differences. In addition, ordinal data frequently used in social sciences and medical studies are not additive in strict sense since equidistant property is not satisfied (Jamieson, 2004). Hand, (1996) opined that for ordinal scale, $\bar{X} > \bar{Y}$ or $\bar{X} < \bar{Y}$ is meaningless. Non-admissibility of meaningful addition may distort standard deviation (SD), correlation, Cronbach α for test reliability etc. Equal importance assigned to the items for summative Likert score ignores different contributions of items to total score, different item-total correlations, different factor loadings, etc. (Parkin et al. 2010). A questionnaire usually has several scales (battery of Likert scales) where the scales differ in terms of number of items (length) and number

of response-categories (width). Here, joint distribution of scale scores is problematic without knowledge of distributions of scale scores. One solution is to convert scores of each item to follow normal distribution where parameters of the distribution can be obtained from the data (Chakrabartty, et al. 2023)

7. Conclusion and Future Work

The proposed transformation using G-inverse is a novel method of introducing perfect correlation between two variables, even if they are non-linearly related. The proposed approaches also avoid the problems of insignificant beta coefficients in univariate and multivariate regression models and presence of outliers. Extension of the concept to perfect linearity in multiple linear regressions and canonical regression will go a long way in empirical researches in various branches of science.

Future studies may be undertaken to find (i) distribution of the proposed perfect correlations after converting item scores of ordinal or categorical data to follow normal distribution and test significance of such correlations along with investigation of effect of each independent variable, and (ii) to compare the efficacy of our suggested approach against other traditional ones by providing quantitative evidences.

References

Agresti A. (2002). *Categorical data analysis (2nd ed).* Hoboken, NJ: Wiley

Bignardi G., Dalmaijer E.S., Astle D.E. (2022): Testing the specificity of environmental risk factors for developmental outcomes. *Child Dev.* **93**:e282–e298. doi: 10.1111/cdev.13719

Brooks, Thomas, Pope, D. and Marcolini, Michael. (2014): Airfoil Self-Noise. UCI Machine Learning Repository. https://doi.org/10.24432/C5VW2C.

Brossart, D. F., Parker, R. I., & Castillo, L. G. (2011). Robust regression for single-case data analysis: How can it help? *Behavior Research Methods, 43*(3), 710–719. https://doi.org/10.3758/s13428-011-0079-7

Box, G. E. P. and Cox, D. R. (1964): An analysis of transformations, *Journal of the RoyalStatistical Society, Series B*, 26, 211-252.

Chakrabartty, Satyendra Nath (2023): Improving Linearity in Health ScienceInvestigations. *Health Sci J.* Vol. 17 No. 4: 1010. DOI: 10.36648/1791-809X.17.4.1010

Chakrabartty, S. N., Kangrui, Wang and Chakrabarty, Dalia (2024): Reliable Uncertainties of Tests & Surveys - a Data-driven Approach. *International Journal of Metrology and Quality Engineering (IJMQE)*.15, 4, 1 – 14. https://doi.org/10.1051/ijmqe/2023018

Cox DR.(1972). Regression models and life-tables (with discussion). *J R STAT SOC* ; B. **34**:187-220. doi: http://dx.doi.org/10.2307/2985181

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Feng, Ge, Peng, Jing, TU, Dongke, Zheng, Julia Z. and Feng, Changyong (2016). Two Paradoxes in Linear Regression Analysis. *Shanghai Archives of Psychiatry*, Vol. 28, No. 6, 355 – 360. https://doi.org/10.11919/j.issn.1002-0829.216084

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy, 98*(Supp. C), 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Fox, S. and Hammond, S. (2017). Investigating the multivariate relationship between impulsivity and psychopathy using canonical correlation analysis. *Personality and Individual Differences*, 111, 187-192. doi:10.1016/j.paid.2017.02.025

Gavurova B., Rigelsky M., Ivankova V. (2020): Perceived health status and economic growth in terms of gender-oriented inequalities in the OECD countries. *Economics and Sociology*, **13**:245–257. doi: 10.14254/2071-789X.2020/13-2/16.

Hand, D. J. (1996): Statistics and the Theory of Measurement, *J. R. Statist. Soc. A*; 159, Part 3, 445-492

Jamieson, S. (2004): Likert scales: How to (ab) use them. *Medical Education,* 38, 1212 -1218

Kim, Y., Kim, T.-H., & Ergun, T. (2015). The instability of the Pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters, 13*, 243–257. https://doi.org/10.1016/j.frl.2014.12.005

Kovacevic, M. (2011): *Review of HDI Critiques and Potential Improvements*, The Human Development Research Paper (HDRP) Series, Research Paper 2010/33.

Liu Y, Ruan J, Wan C, Tan J, Wu B, Zhao Z. (2022): Canonical correlation analysis of factors that influence quality of life among patients with chronic obstructive pulmonary disease based on QLICD-COPD (V2.0). *BMJ Open Respir Res*. 9(1):e001192. doi: 10.1136/bmjresp-2021-001192.

Loco, J.V; Elskens, M., Croux, C. and Beernaert, H. (2002). Linearity of calibration curves: use and misuse of the correlation coefficient. *Accreditation and Quality Assurance* (7):281–285. DOI 10.1007/s00769-002-0487-6

Malakar B., Roy S.K., Pal B. (2022): Relationship between physical strength measurements and anthropometric variables: Multivariate analysis. *J. Public Health Dev.* **20**:132–145. doi: 10.55131/jphd/2022/200111

Mardia, K.V. and Bibby, J.M. and Kent, J.T. (1982): *Multivariate analysis*, Academic Press

Niven, E. B., & Deutsch, C. V. (2012). Calculating a robust correlation coefficient and quantifying its uncertainty. *Computers & Geosciences, 40*, 1–9. https://doi.org/10.1016/j.cageo.2011.06.021

Parkin D, Rice N, Devlin N.(2010): Statistical analysis of EQ-5D profiles: does the use value sets bias inferences? *Med Decis Making* 30(5): 556–565. DOI: 10.1177/0272989X09357473

Rao, C. Radhakrishna and Mitra, Sujit Kumar (1971). *Generalized Inverse of Matrices and its Applications*. New York: John Wiley & Sons. *ISBN 978-0-471-70821-6*

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633–639. https://doi.org/10.1080/01621459.1990.10474920

Song-Gui Wang & Shein-Chung Chow (1987): Some results on canonical correlations and measures of multivariate association. *Communications in Statistics - Theory and Methods*, 16:2, 339-351, DOI: 10.1080/03610928708829370

Stefano, Claudio; Fontanella, Francesco; Maniaci, Marilena and Freca, Alessandra (2018). Avila. UCI Machine Learning Repository. https://doi.org/10.24432/C5K02X

Vasylieva T, Gavurova B, Dotsenko T, Bilan S, Strzelec M, Khouri S. (2023): The Behavioral and Social Dimension of the Public Health System of European Countries: Descriptive, Canonical, and Factor Analysis. *Int J Environ Res Public Health*. 20(5):4419. doi: 10.3390/ijerph20054419.

Wessa P. (2012): Box-Cox Linearity Plot (v1.0.5) in Free Statistics Software (v1.1.23-r7),Office for Research Development and Education. http://www.wessa.net/rwasp_boxcoxlin.wasp/

Wilcox, R. R. (2023). Robust Correlation Coefficients That Deal With Bad Leverage Points. *Methodology*, Vol. 19(4), 348–364. https://doi.org/10.5964/meth.11045

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic Press.

Yellowlees, A., Bursa, F., Fleetwood, K. J., Charlton, S., Hirst, K. J., Sun, R., & Fusco, P. C. (2016). The appropriateness of robust regression in addressing outliers in an anthrax vaccine potency test. *Bioscience, 66*(1), 63–72. https://doi.org/10.1093/biosci/biv159