Avoiding Mistakes in Bivariate Linear Regression and Correlation Analysis, in Rigorous Research

László Barna Iantovics

George Emil Palade University of Medicine, Pharmacy, Science and Technology of Targu Mures, Gheorghe Marinescu, 38, 540142, Targu Mures, Romania, barna.iantovics@umfst.ro; ORCID iD: https://orcid.org/0000-0001-6254-9291

Abstract: Data science and artificial intelligence are emergently, very fast-evolving fields, being applied to a large diversity of real-life problem-solving. In this context, some methods are applied without verifying assumptions that must be met, for the correct applicability and the necessary model fit. Such mistakes could lead to misinterpretations of the results. One of the application domains, that is very affected in this sense, is healthcare, where misinterpretations could have dangerous effects on human health. Based on an in-depth study of the scientific literature, it was identified that bivariate linear regression (BLR) even is considered simple, is one of the methods that sometimes leads to confusion in application. With this in mind, this paper proposes in an algorithmic form of a methodology that consists of assumptions, that must be passed by the BLR, so that the applicability is correct and should pass the required threshold model fit. Also, presented in algorithmic form is the decision for the correct calculus of the bivariate linear correlation coefficient (BCC). There are other considerations, like the necessary sample sizes for the two variables in the case of BCC and BLR. The proposed methodology, herein, will be useful for researchers, since BLR is frequently applied in research in diverse domains, like industry and healthcare, individually or combined with methods of data science and artificial intelligence.

Keywords: data science; linear regression; model fit; prediction; artificial intelligence; mathematical modeling; goodness-of-fit; mistakes encountered in clinical research; correlation coefficient; data misinterpretation

1 Introduction

Methods of Data Science and Artificial Intelligence including Intelligent Systems are successfully applied for a wide diversity of real-life problem-solving. An optimization method of robotic mobile agent navigation uses a neural network [1]. In [2] a comprehensive review of recent trends in measuring machine intelligence is presented. Even though measuring machine intelligence is of high interest, there are very few methods focused on measuring machine intelligence [3] [4]. Methods based on statistics combined with methods of artificial intelligence and data science are frequently applied together to combine the advantages of the constituting methods. A novel statistical methodology is applied for the detection of cooperative multiagent systems with extreme intelligence (those that are statistically significantly more intelligent than a set of considered intelligent systems) [5].

Research in all domains, industry, and healthcare, comprises problems or subproblems that involve bivariate correlation (BC) and/or regression analysis (RA) which includes the bivariate linear regression (BLR) analysis [6]. For instance, soil water content prediction using regression models could offer support in the decision-making processes [7]. The research [8] focused on the study of the bivariate correlation between health literacy and cell phone addiction among Iranian healthcare students. A frequent mistake in research that involves BC and BLR is that these methods are applied without verifying the necessary assumptions, which should be passed for model fit and correctness of their applicability even in papers published in the best journals. This could lead to erroneous interpretation of the research results. One of the most affected domains in this sense is healthcare, where misinterpretations of BC and BLR could lead even to loss of lives [9]. Based on this motivation, this paper proposes a mathematically grounded modeling of the assumptions that must be passed by BLR to be applicable and to pass the requested threshold model fit. At the same time, it presents the algorithmic decision for the correct calculus of BCC.

The upcoming structure of the paper is organized as follows: Section 2 presents a brief survey of the state-of-the-art research that is based on BC and BLR; Section 3 presents the proposed methodology; and Section 4 presents the experimental evaluation of the proposal. Finally, in Section 5, the conclusions of the paper are presented.

2 State-of-the-art Applications of BC and BLR

BC and BLR have applications in many real-life problems solving. Prediction could be helpful in human decision-making. Frequently prediction methods are based on bivariate linear regression. Prediction based on BLR is approached in various research, such as rice sheath blight field resistance prediction [10], reduced energy consumption prediction [11], and many others.

Anti-social behavior identification in online discussions frequently is important to be identified. For this problem solving a classification method that involves diverse regression methods is proposed [12].

Concerns of decision-makers toward the profit obtained by using cloud computing technology are studied in [13]. The proposed method includes linear regression.

A study on the suitability of predictive control on an advanced mechatronic system consisting of a laboratory helicopter is presented in [14]. The developed model includes a linear regression method.

Sometimes linear regression is combined with diverse methods of artificial intelligence or data science for solving prediction problems like traffic with climate condition prediction [15], city-wide demand-side prediction [16], and many others.

Various regression methods are applied for diverse real-life problem-solving. Another widely applied regression method is logistic regression, applied for problem-solving like identifying candidate disease genes [17], sampling on-demand [18], semantic web service matchmaking [19], etc.

Obtaining real data in industry and healthcare for research frequently is difficult. A novel method for data quality assessment of synthetic data obtained by simulation is based on a statistical approach [20]. It is treated the bivariate logistic regression that is frequently applied mostly in healthcare-related research. There are proposed mathematically grounded assumptions that must be met for the application of bivariate logistic regression to be correct and pass the requested threshold model fit.

The presented bibliographic study shows that even though BC and BLR are traditional methods, they still have even recent times many applications.

3 Assumptions Testing in BC and BLR

3.1 Mistakes in Bivariate Correlation and Regression Analyses

Even simple descriptive statistics and data normality analyses presented in research have some common mistakes [21]. A decision rule for data central tendency indicator establishment in [22] is presented. In [23] is treated the subject of avoiding mistakes in quantitative statistical analyses in political science. A valuable step-by-step guide for the correct application of different statistical tests is presented in [24].

There are some common statistical errors that appear in papers published in radiology journals [25]. The study involved 157 selected papers from 20 radiology journals that were published between 2016 and 2017. The selected articles were assessed regarding different kinds of statistical errors like mistakes in statistical tests applied, wrong interpretation of p values, and some others. The mistakes were treated considering the journals based on their impact factors.

According to [26], the most common mistakes in some medical research consist of wrong sample size determination; mistakes in bias related to sampling; mistakes in making adjustments in multiple comparisons; wrong interpretation of the p-value by considering clinical relevance, choosing wrong statistical tests, etc.

A guide for avoiding some mistakes in scientific investigations related to epidemiology, and public health in [27] is presented. There are suggested questions that must be responded to before the effective beginning of a research.

In many papers, there are reported mistakes that could appear in the statistical regression analyses [28-31]. Statistical analyses performed on animal science present some common mistakes that are presented in [32]. Studies on confounding bias for heritability include different mistakes that are treated in [33]. A study on the biases in summary statistics of slopes and intercepts in linear regression that includes errors in both independent and dependent variables is presented in [34].

As a general conclusion based on the performed bibliographic study can be formulated that the usual mistake in the case of statistical methods based on linear regression is that they are applied without being based on necessary assumptions that should be met for their applicability to be correct, and the model fit to be passed the required threshold. This laziness frequently leads to misinterpretations. Linear correlation and regression analysis are among these methods.

3.2 The Proposed Validation and Analysis Methodology

In the following, are presented the assumptions that must be passed by BLR and the model fit analyses. Also, the correct calculus of BCC is treated. The presented mathematically grounded assumptions are applicable in any type of research that involves BC and BLR.

The BLR problem includes two variables measured at continuous levels (interval or ratio variables) denoted in the following *VrX* and *VrY*. $|VrX|$ and $|VrY|$ denote the sample sizes of *VrX* and *VrY*. $|VrX|=|VrY|=n$, $Df=n\text{-}2$, where *Df* represents the degrees of freedom. The *BivRegMet* algorithm presents the methodology for verification of the assumptions that should be met for BLR could be applicable and passing the necessary model fit thresholds. Previously to *BivRegMet* could be applied the *BivCorr* algorithm that describes the correct calculus of the correlation coefficient of *VrX* and *VrY*. PCc denotes the Pearson correlation coefficient. SCc denotes the Spearman correlation coefficient. In case *BivCorr* is applied, then the *BivRegMet* algorithm application can be decided on the response to the fact that it is a valuable parametric or nonparametric statistics, being applicable only in parametric case (both *VrX* and *VrY* passed the normality assumption). However, *BivCorr* if applied is a prefiltering assumption and model fit verification for the BLR application.

It must be noticed that in many types of research, the human evaluator (*HE*) should have a central role in the interpretation of the experimental evaluation results and the establishment of the values of different parameters. According to the two algorithms, *HE* is responsible for the establishment of the parameter's values (*CL*, α_{anova}, etc.), thresholds for model fit (necessary for the interpretation; the threshold for the strength of the correlation for instance), and supervises the evaluation process, by deciding in different decision points when necessary. Must be motivated that *HE* must make some visual examinations of graphical results. *HE* in his/her contribution will consider the specificity of the research, application area, and his/her background knowledge.

BivCorr: Bivariate Correlation Coefficient Calculus algorithm

```
IN: VrX; VrY; OUT: r; //correlation coefficient.
SignIndic; CorrStr;// significant correlation is detected?; correlation strength
Direc; // correlation direction: positive or negative?;
Step A1. Calculus of the correlation coefficient.
CL:=95%; //the implicit confidence level
Step A1.1. Verification of the normality assumptions.
@Verification of VrX and VrY normality using numeric goodness-of-fit tests;
@VrX, VrY normality results validation by visual examination of the Q-Q plots;
Step A1.2. Calculus of the correlation coefficient.
If (VrX and VrY are normally distributed) Then @Calculates PCc r. //parametric
    Else @Calculates SCc r. //nonparametric. $\rho$ is more usual notation.
EndIf
If ($r = 0$) Then @ no correlation, the analysis is stopped. EndIf
Step A2. Assessment of correlation significance and correlation direction
@Establish the Research Hypotheses:
    $H_{cr}$: "r is statistically equal with 0" //null hypothesis, no correlation
    $H_{cra}$: "r is statistically significantly different from 0" //alternative hypothesis
@Calculates the CI, [Lr, Ur] of r at the CL confidence level.
If  (($r > 0$) and ($Lr > 0$))  Then //significant positive correlation.
  $H_{cr}$ rejected, $H_{cra}$ accepted; SignIndic := "Yes"; Direc := "+";
  ElseIf  (($r < 0$) and ($Ur < 0$)) Then //significant negative correlation.
      $H_{cr}$ rejected, $H_{cra}$ accepted; SignIndic := "Yes"; Direc :="-";
        Else $H_{cr}$ accepted. @there is no correlation, the analysis is stopped.
EndIf
Step A3. Establishment of the correlation strength.
```

```
If (SignIndic="Yes") Then @The strength of the correlation CorrStr is
    established based on the |r| value considering the classification from Table 1.
EndIf
EndBivCorr
```

In the case of both algorithms, each of the parameters is set to an implicit value that usually is considered the most appropriate. *HE* could change these values if considered, based on his/her consideration and taking into account other evidence if available (for instance in similar research, is obtained, a certain value of the correlation coefficient and in the actual study, is considered to have a comparable value).

α denotes the Type I error rate. It is recommended to approach the two-tailed α. β denotes the Type II error rate. The power is calculated as 1-β. r_{estim} represents an estimation of the correlation coefficient (the expected correlation coefficient). r_{estim} value can be based on some background knowledge (previous study for instance). Z_α denotes the standard normal deviate for α. Z_β denotes the standard normal deviate for β. It is given, β and an estimate of expectable r_{estim} size then can be calculated the necessary sample size *n* (and degree of freedom *Df*) (1, 2) [35]. For example, α=0.05 and power=0.8 (β=0.2), when correlation coefficients are increasing 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9, the sample sizes are decreasing as follows 193, 84, 46, 29, 19, 13, 9 and 6, respectively.

$$C = 0.5 \times \ln[(1+ r_{estim})/(1- r_{estim})] \tag{1}$$

$$n = [(Z_\alpha+Z_\beta)/C]^2 + 3 \tag{2}$$

Step A1.1, in the case of numerical evaluation of the normality assumption, for small sample sizes (n≤30), as a decision rule, recommended the application of the Shapiro-Wilk (SW) goodness-of-fit test [36], which is proven [37] as having the highest statistical power compared with the frequently used: Kolmogorov-Smirnov, Lilliefors (Lill), and Anderson-Darling tests. A limitation of the SW test consists of the sensitivity in the case of large samples. In case of larger sample sizes, we recommend the application of the Lill test [38-40] which represents an adaptation of the Kolmogorov-Smirnov test [41].

The Quantile-Quantile (Q-Q) plot [42] [43] is a scatter plot appropriate for the normality visual validation. A Q-Q plot is a drawn reference line. The visual study of the data normally involves the verification if the data points fall almost along this reference line. The larger the difference from the reference line, the larger the evidence is for the interpretation that the data fails to pass the normality assumption. Additionally, to the mentioned numerical verification of normality, it is recommended to make a visual validation based on the drawn Q-Q plot.

r denotes the correlation coefficient of *VrX* and *VrY*. According to Step A1.2 of the algorithm, if both *VrX* and *VrY* are normally distributed, the PCc *r* [44] [45] is computed, elsewhere the SCc *r* (more precisely notation is ρ) [44] [45] is computed. This decision, regarding the calculus of the correlation coefficient, is based on the fact that SCc is more robust than PCc (less sensitive to influential points), and considering this circumstance, it is more appropriate in the nonparametric case. As additional validation, *HE* can make a visual evaluation of the Scatter Plot created based on *VrX* and *VrY*. This simple approach is useful to visually present the relationship between the two studied continuous variables; indicate if there are influential points; and estimate if the relationship is linear.

When r=0, this indicates no correlation. Step A2 verifies if the correlation is statistically significant (if r is statistically significantly different from 0) based on the r value and the *CI* of r at the *CL*% level. If the difference is statistically significant, and if $r>0$ then the correlation is positive, if $r<0$ the correlation is negative.

Step A3 establishes *CorrStr* as the strength of the correlation according to Table 1 in case a statistically significant correlation is detected in the previous step. *CorrStr* value can be considered a model fit measure. *HE* could require a certain level of correlation strength. For instance, could establish that just a very strong correlation is acceptable.

Table 1

Range of correlations strength

Correlation coefficient value	**Interpretation (Level of correlation strength)**
$\lvert r\rvert \in [0.8, 1]$	Very strong
$\lvert r\rvert \in [0.6, 0.8)$	Strong
$\lvert r\rvert \in [0.4, 0.6)$	Moderate
$\lvert r\rvert \in [0.2, 0.4)$	Week
$\lvert r\rvert \in [0, 2)$	Neutral

Additionally, in the case of parametric correlation is recommended the calculus of the r^2, $r^2 \in [0,1]$ is called the coefficient of determination [46]. r^2 is an indicator of the effect size. r^2 is the proportion of the variation in the *VrY* variable that is explainable/predictable by the *VrX* variable. For instance, r^2=0.82 indicates that 82% of the variance of the *VrY* variable is explained by the variance of the *VrX* variable. r^2 is a measure of the goodness-of-fit of the model, higher value means better model fit. The minimal required threshold value of *CD* that should pass r^2 must be established by *HE* based on the specificity of the research and considering how good the model fit should be. Frequently *CD*=0.7 can be considered as an implicit parameter value. In case if $r^2 \geq CD$ then the threshold passed, else the threshold does not pass. Table 2 presents the usual interpretation of r^2 values. *HE* will consider also the strength of the correlation jointly with the value of r^2 (in the case of parametric correlation is interpreted r^2).

Table 2

r^2 interpretation in parametric case

r^2	Interpretation
≥0.85	Very good
[0.75, 0.85)	Good
[0.6, 0.75)	Satisfactory
<0.6	Weak

Spearman's *r* (more precisely *ρ*) is an indicator of monotonicity. It reflects the extent to which an increase/decrease in *VrX* is associated with an increase/decrease in *VrY*, but the expanse of increase/decrease does not have to be constant over the whole range of values, as in the case of linear correlation.

BivRegMet: Assumptions for application of BLR

```
IN: VrX; VrY; Principal OUT: $y(x)$;//regression equation
The visual plotted regression line and its confidence interval (CI);
Secondary OUT: DW// the Durbin-Watson statistics result.
SignSlope;//slope of the regression line is significantly different from 0;
RD;//root mean square deviation (RMSD), called standard deviation of residuals;
Step B1. Preliminary assumptions checking
@Obtain the DW applying the Durbin-Watson test.
@VIS1:The residual plot is created, plotted standardized predicted scores (X axes) against standardized residuals (Y axes). HE visually verifies the homoscedasticity.
@VIS2:Elaborated probability–probability (P-P) plot. HE makes a visual examination of the normality of residuals.
@HE establishes $R_{MSD}$ value. Calculate RD and the value of RMSD.//model fit
Step B2. Model Fit Overall Model test
@Establish the Research Hypotheses:
  $H_r$: The slope (denoted $a$) of the regression equation is statistically equal to 0.
  $H_{ra}$: The slope of the regression equation is significantly different from 0.
@Apply the ANOVA test. Let $Pval_{an}$ be the obtained p-value of the ANOVA test.
If ($Pval_{an}>\alpha_{an}$) Then $H_r$ is proved. SignSlope:="No"; //NO significant difference.
  Else $H_r$ proving failed. $H_{ra}$ proved. SignSlope:="Yes";//significant difference.
EndIf
Step B3. BLR modeling if assumptions passed and model fit
If (($DW$~2) and (VIS1 passed) and (VIS2 passed) and ($RD<R_{MSD}$) and (SignSlope="Yes")) Then
    @Constuct the linear regression equation $y(x) = a \times x + b$.
    @Plot the regression line including the CI at the CL level.
    @HE makes a visual validation of the regression line considering also its CI.
    @HE treats influential points if exist.//stepwise methodology described in
    Section 4 in applicative form.
        Else @"weak model-fit".//BLR is not approachable
EndIF
EndBivRegMet
```

Step B1 consists of some preliminary analyses. For the verification of the independence of values in the two variables (autocorrelation in the residuals) is applied the Durbin-Watson statistic obtaining a *DW* value. $DW\in[0,4]$, where $DW=2$ indicates there is no autocorrelation, $DW<2$ indicates positive autocorrelation, and $DW>2$ indicates negative autocorrelation. The residuals must have a constant variance with no dependence on the level of the dependent variable, a property called homoscedasticity (VIS1 verification). After the Durbin-Watson test, *HE* performs the visual examination of homoscedasticity on the residual plot created verifying that the error terms variance is constant crossways the dependent variable values. *HE* makes a visual examination of residuals normality on the P-P plot (VIS2). RMSD is a measure of the goodness-of-fit of the regression line. *RD* represents the calculated RMSD value. *RD* must be interpreted according to Table 3, comparatively with an established threshold value $\mathrm{R}_{\mathrm{MSD}}$.

Step B2 responds to the Null Hypothesis *Hr* and Alternative Hypothesis *Hra*, for whose verification is applied the ANOVA test. $Pval_{an}$ denotes the obtained p-value of the ANOVA test applied at the α_{an} significance level. α_{an} could have different values like 0.01, 0.001, etc. In most cases is recommended the α_{an} value 0.05. In Step B2 performed the regression analysis, where *HE* will decide on the application of BLR, based on the requirements DW~2; VIS1 visual examination passing; VIS2 visual examination passing; $RD<R_{MSD}$; and *SignSlope*="Yes".

Table 3

Root Mean Square Deviation Interpretation

RMSD value	Interpretation
≤ 0.75	Very good
(0.75, 1]	Good
(1, 2]	Satisfactory
>2	Not satisfactory

Finally, the regression equation is constructed and the regression line is plotted with the CI at the CL level being appropriate for visual validation of model fit. The CI helps in the visual and numerical appreciation of points that fall outside the CI. *HE* will use this information in making the final decision on the effective application of bivariate linear regression for the considered problem-solving. Influential points are important to be detected in regression analysis since they could have a large impact on the linear regression equation [47] [48]. Section 4 additionally presents applications of the step-by-step treating of influential points that could largely affect the regression equation.

4 Experimental Evaluation

In this section, for testing and evaluation of the proposed methodology it was performed an experimental evaluation case study focused on a prediction problem based on BLR. In this case, *VrX* is called independent (predictor) variable and *VrY* is called dependent (predicted) variable. Also, we present some additional elements that could be used by *HE* in the interpretation of the results and formulation of conclusions based on data analysis results. Table 4 presents the generated synthetic data used in the evaluation. The subscripts labeled "inf1", "inf2" and "inf3" indicate influential points that will be identified and treated later.

Table 4

The data used for experimental evaluation

VrX	***VrY***	***VrX***	***VrY***	***VrX***	***VrY***
93	3.78	64	2.88	71	2.89
$61^{\mathbf{inf1}}$	$3.8^{\mathbf{inf1}}$	$62^{\mathbf{inf2}}$	$1.6^{\mathbf{inf2}}$	$73^{\mathbf{inf3}}$	$2.42^{\mathbf{inf3}}$
74	3.1	85	3.19	87	3.44
69	2.88	94	3.68	91	3.91
70	3.21	78	3.28	56	2
53	2.1	66	3.1		

Initially is applied *BivCorr* algorithm. Step A1.1, considering the low sample size of both variables $|VrX|=|VrY|=17$ (17<30), $Df=15$, the best option is the application of the SW test and making a visual validation of normality based on the Q-Q plot. For illustrative purposes of the interpretation of diverse results that could be obtained by different normality goodness-of-fit tests, the Lill test, along with the SW test was applied. For both tests, it was considered the $\alpha_{norm}=0.05$ significance level. Table 5 presents the results of the Lill and SW tests applied to both variables *VrX* and *VrY*. Both variables met the normality assumption for both tests of normality. *p* denotes the p-value obtained after application of a test, with $p>\alpha_{norm}$, which leads to the acceptance of the null hypothesis of the normality test, (H_0) that states the assumption of normality.

Table 5

The results of the applied normality tests

	Lilliefors			**Shapiro-Wilk**		
Variable	***Statistic***	***P***	$p > \alpha_{norm}$	***Statistic***	***P***	$p > \alpha_{norm}$
VrX	0.127	0.2	Yes	0.949	0.438	Yes
VrY	0.185	0.126	Yes	0.936	0.271	Yes

Figures 1 and 2 present the Q-Q plots for *VrX* and *VrY*. The visual interpretation of both figures leads to the same conclusions as the numerical SW and Lill tests for meeting the normality assumption.

In the following, a descriptive static (Table 6) was performed, even if it is not integrated into the algorithms. This could admit the formulation of some additional remarks. *SD* denotes the standard deviation. *Variance* represents the variance that is calculated as SD^2. For each variable, the mean was chosen as the central tendency indicator since both variables met the normality assumption. *CV* denotes the Coefficient of Variation, $CV=SD/mean\times 100$. The lower the *CV*, the lower the dispersion. *CV* admits the evaluation of data homogeneity-heterogeneity, by making a classification: $CV\in[0, CVa)$ indicates homogeneity; $CV\in[CVa, CVb)$ indicates relative homogeneity; $CV\in[CVb, CVc)$ indicates relative heterogeneity; $CV\geq CVc$ indicates heterogeneity. The most usual recommended values, which should be established by *HE*, are CVa=10, CVb=20, and CVc=30.

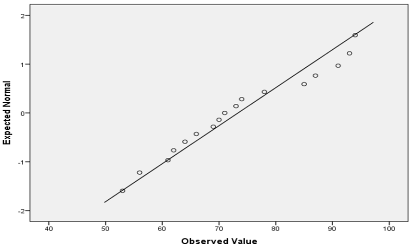

Figure 1

The Q-Q plot of *VrX*

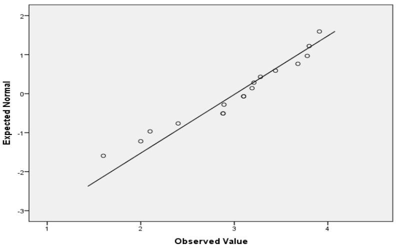

Figure 2

The Q-Q plot of *VrY*

Table 6

Descriptive statistics

Variable	*Mean*	*SD*	*Variance*	*CV*
VrX	73.353	12.85	165.123	17.52
VrY	3.014	0.665	0.442	22.06

Table 7 presents the results of descriptive statistics obtained by performing bootstrapping based on 1000 samples. L_{rd} denotes the lower bound of the 95% CI. L_{ur} denotes the upper bound of the 95% CI. *SE* denotes the Standard Error.

Table 7

Descriptive statistics results by bootstrapping

Variable		*Bias*	*SE*	L_{rd}	L_{ur}
VrX	Mean	-0.025	3.102	67.237	79.529
	SD	-0.628	1.622	8.832	15.095
VrY	Mean	0.005	0.155	2.695	3.311
	SD	-0.035	-0.035	0.408	0.825

According to Step A1.2, since both variables meet the normality assumption it was chosen the calculus of *r* as the PCc, with r=0.712 (Table 8). r>0 indicates the possibility of the existence of a positive linear correlation. In the column labeled

"*" the CI of r is presented at the CL=95%. $r>0$ and 0.351>0 indicate that there is a statistically significant positive correlation. In the column labeled "**" the 95% CI of r is calculated based on bootstrapping, with Bias=0.002 and SE=0.154, r proving to be statistically significant even at the 0.01 level. Bootstrapping was applied using 1000 samples. $r>0$ and 0.358>0 led to the formulation of the same conclusion, claiming the existence of a positive correlation.

Table 8

Results of correlation analysis

Pearson *r*	***p-value***	r^2	**95% CI of *r* ***	**95% CI of *r* ****
0.712	0.001	0.51	[0.351, 0.888]	[0.358, 0.935]

Step A3, According to the classification presented in Table 1, r=0.712 ($|r|\in[0.6, 0.8)$), indicate a strong linear correlation. The obtained p-value 0.001 indicates that the correlation is significant even at the 0.001 level. Since it is applied parametric statistics it is calculated the coefficient of determination r^2. *HE* set the *CD* value to 0.7. r^2, r^2=0.51, 0.51< *CD* indicates that the passing of the threshold *CD* failed, and even if the linear correlation is strong the model is not appropriate.

Anyway, *HE* decided on the continuation of the analysis also considering the existence of potential influential points. As a second step, based on the parametric statistics the BivRegMet algorithm was applied. It was obtained DW=1.73 very close to value 2, indicating slight negative autocorrelation. Figure 3 shows heteroscedasticity which is a violation of an assumption that should be passed (VIS1). Visual examination of Figure 4 shows the violation of the residuals normality assumption, the residuals of the regression do not follow a normal distribution (VIS2). The *RD* value 0.4824, according to Table 3 admits the formulation of very good goodness-of-fit of the regression line. These analysis results indicate the violation of some assumptions that should be passed, based on this fact *HE* could conclude that the model in the actual form is not appropriate, eventually if there are influential points their removal could have a remediating effect. In the following, there were formulated the hypotheses *Hr* and *Hra*. It was applied the ANOVA test, at the significance level α_{an}=0.05, with results in Table 9, with the predicted variable *VrY* and the predictor variable *VrX*. $Pval_{an}<\alpha_{an}$ admits rejection of *Hr* and acceptance of *Hra* according to that the slope of the regression equation is statistically different from 0.

Table 9

Results of the ANOVA test, indicator of the Overall Model Test

Model	**Sum of Squares**	**Df**	**Mean Square**	**F**	$Pval_{an}$	p>α_{an}
Regression	3.588	1	3.588	15.422	0.00134	No
Residual	3.49	15	0.233			
Total	7.079	16				

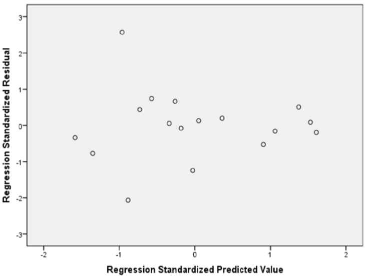

Figure 3
Scatterplot for visual evaluation of Homoscadacity: VaY

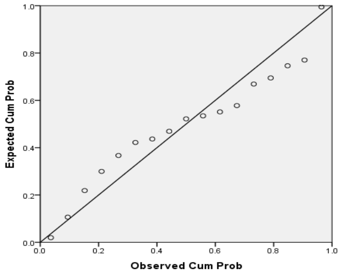

Figure 4
Normal P-P Plot of Regression Standardized residual: Dependent variable VaY

Table 10 presents the coefficients of the regression equation and for additional appreciation the SE, the lower bound (Lb), and the upper bound (Ub) of the 95% CI.

The obtained regression equation (3) is:

$$y(x) = 0.311 + 0.037 \times x \tag{3}$$

Table 10
The coefficients of the regression equation

Model	**Best-fit value**	**SE**	**Lb**	**Ub**
Slope	0.037	0.009	0.017	0.057
Y Intercept	0.311	0.698	-1.178	1.799
X Intercept	-8.431			

Figure 5 presents the plotted linear regression line, with the 95% CI. The visual interpretation of Figure 5 shows that 29.4% of points fall outside the 95% CI. This leads to the formulation of the remark of indication of weak prediction power.

Based on the previous analysis *HE* should formulate the conclusion that in this case, the model fit is above expectations of *HE* in performing prediction based on bivariate linear regression and another method should be chosen if all the actual data is available.

α is the probability of a Type I Error. β is the probability of a Type II Error. *power* denotes the power. Additionally, it was performed a post-hoc analysis by computing the achieved power, based on considered two-tails, r=0.712, the sample size=17, and established significance level α=0.05 obtaining the *power*=0.931, where β=0.069 (*power=1-β*).

HE based on the visual interpretation of the graphical representation of the regression line and its CI, identifies (X,Y)=(61,3.8) (marked in Table 4 with "inf1") as a potential influential point that has a large influence on the regression line. HE decided on (61,3.8) removal and the analysis has been repeated. As result obtained: *r*=0.853 (increased with 0.141) indicates a very strong correlation; r^2=0.728 (increased with 0.218), with the threshold CD (CD=0.7), where r^2>CD, now this assumption is passed (initially with the "inf1" included does not passed); *RD*=0.3532 (decreased with 0.1292), which indicates a slight improvement; the new *DW* value decreased to 1.274 (that is worse than 1.723), indicating degradation (that is conflictual to the other indicators). Based on these facts *HE* decided on the removal of (X,Y)=(61,3.8) and continuation of the exploratory analysis of assumption passing, increase the model fit and the model prediction power.

HE based on the visual evaluation of the graphical regression line and its CI (Figure 5) identifies (X,Y)=(62,1.6) (marked in Table 4 with "inf2") as a potential influential point. Removing this influential point and repeating the analysis resulted in, an increased *r*=0.887 (additional increase with 0.034) indicating a very strong correlation; r^2=0.787 (additional increase with 0.059), where r^2>*CD* (*CD*=0.7); *RD* value decreased to 0.27 (additional decrease with 0.0832), which indicated a better model fit; the new *DW* value becomes 1.588 which is slightly better than 1.274 but is still worse than the initial 1.723 (with all the influential points included). *HE* decided on (62,1.6) removal and continuation of the exploratory analysis. The obtained regression equation with the two influential points, marked with "inf1" and "inf2" removed is presented in (4), plotted in Figure 6 with the 95% CI.

$$y(x)= 0.152 + 0.0388 \times x. \quad (4)$$

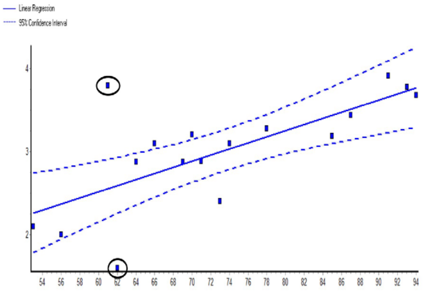

Figure 5

The regression line with the 95% CI

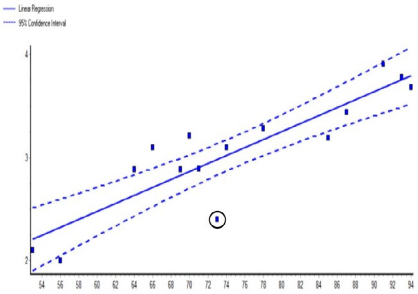

Figure 6

The regression line with the 95% CI with "inf1" and "inf2" influential points removed

HE based on the visual interpretation of the graphical representation regression line and its CI (Figure 6), identifies (X,Y)=(73,2.42) (marked in Table 4 with "inf3") as a potential influential point removing it for performing an exploratory analysis. The obtained regression equation with the three influential points, "inf1", "inf2" and "inf3", removed is (5) plotted in Figure 7 with the 95% CI.

$$y(x)= 0.2324 + 0.03824 \times x \qquad (5)$$

Applying BivCorr resulted in, increased r=0.924, p>0.001(additional increase with 0.037), indicating a very strong correlation (95%CI=[0.771,0.976]); increased r^2=0.853 (additional increase with 0.066), where r^2>CD, with the threshold *CD* passed. Applying BivRegMet, the *DW* value of 1.997~2 indicates no autocorrelation. Figure 8 indicates homoscadacity. Figure 9 shows that the assumption of residuals normality has been met. *RD* value decreased to 0.221 (additional decrease with 0.049).

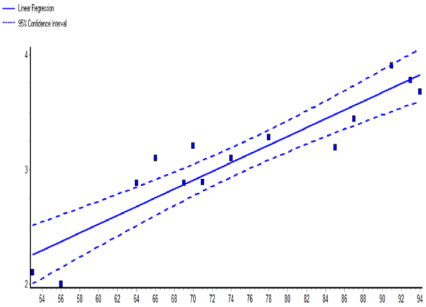

Figure 7

The regression line with the 95% CI with "infl", "inf2" and "inf3" influential points removed

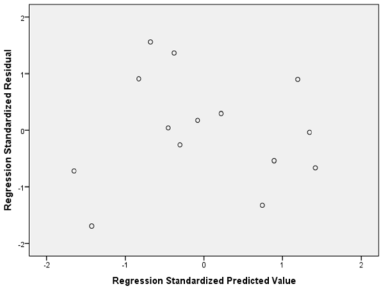

Figure 8

Scatterplot for evaluation of Homoscadacity: VaY with all the influential points removed

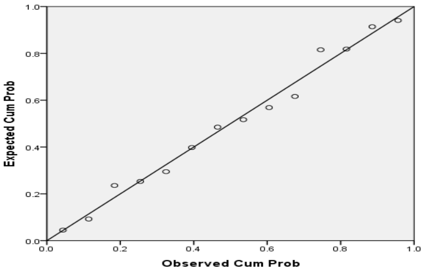

Figure 9

Normal P-P Plot of Regression Standardized residual: VaY with influential points removed

Studying Figure 7 visually, even though a few points fall outside the CI they do not fall too far. To do not make mistakes regarding overfitting can be considered that is not appropriate their removal. The application of the methodology proved that the BLR is applicable and the removal of all the influential points resulted in a better model fit, all the necessary assumptions were met, the final model having a better prediction power. It was proven also the important role of *HE* in evaluation even if some assumptions passed the required threshold the visual interpretation revealed influential points whose removal have improving effect.

Must be noticed that even a very strong correlation between two variables *VarX* and *VarY* does not prove causality. It could happen that *VarX* cases *VarY* or vice versa or there is a third factor *VarZ* that gives growth to the variation in both variables. This is also a situation that shows the important role of *HE*.

Conclusions

In the case of many real-life problems, methods based on statistics, which are sometimes combined with methods of artificial intelligence or data science, are frequently applied. In research that includes bivariate linear regression (BLR), assumptions that must be met for the applicability and the necessary model fit frequently are missed or wrongly applied. Because of this, there is no clear response to the question of whether the BLR is appropriate for a certain problem or subproblem solving. Among others, this is a usual situation in healthcare-research, where misinterpreted data analysis results could have dangerous effects.

With this in mind, this work presented, in the form of algorithmic methodology, the proposed assumptions that must be met for the correct application of BLR and the measurement of the strength of model fit, passing the model fit threshold. It presents an experimental testing and evaluation of the proposed methodology. The proposals from this paper, will be a useful source for researchers who would like to measure the strength of linear correlations, between two variables and/or apply BLR individually or combined with methods of artificial intelligence, data science, or other statistical methods for problem-solving to avoid making mistakes. The methodology can be applied to any type of research that involves BCC and BLR.

It must be said that in many types of research, the *HE* should have a central role in the interpretation of the experimental evaluation results, based on its human-specific background knowledge and contextual knowledge concerning the solved problem. For instance, this can be considered the case of prognosis based on time series, when the data collected in case of a phenomenon meets all the assumptions but is not characteristic of the phenomenon the linearity. For example, we consider the time series of a currency (dollar, euro) that could have a linear tendency over some time but the common sense of *HE* could indicate that linear regression-based prognosis is not a good approach. A proof regarding the importance of the *HE* role in our methodology consisted of the visual discovery of the influential point whose further elimination has as a result fulfilled all the necessary (numeric and visual) assumptions for applicability and model fit threshold. HE applied a trial-and-error methodology when testing the influence of the supposed influential points. For the autocorrelation tests the removal of the first two influential points had a negative effect but after the removal of the third influential point the situation was remediated *DW* indicated no autocorrelation. *HE* in the decisions that make should consider the specificity of the research, application area, personal experiences/knowledge and different background knowledge (other studies results).

Acknowledgment

This work is a part of the CHIST-ERA program supported by the Future and Emerging Technologies (FET) program of the European Union through the ERA-NET funding scheme under the grant agreements, title Social Network of Machines (SOON). This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CCCDI-UEFISCDI, project number 101/2019, COFUND-CHIST- ERA-SOON, within PNCDI III. We would like to thank you for the Research Center on Artificial Intelligence, Data Science, and Smart Engineering (ARTEMIS).

References

[1] Brassai, S. T., Iantovics, L. B., Enachescu, C., Optimization of Robotic Mobile Agent Navigation, Studies in Informatics and Control, 21(4), 2012, pp. 403-412

[2] Iantovics, L. B., Gligor, A., Niazi, M. A., Biro, A. I., Szilagyi, S. M., Tokody, D.: Review of Recent Trends in Measuring the Computing Systems Intelligence, BRAIN - Broad Research in Artificial Intelligence and Neuroscience, 9(2), 2018, pp. 77-94

[3] Iantovics, L. B., Rotar, C., Niazi, M. A. MetrIntPair-A Novel Accurate Metric for the Comparison of Two Cooperative Multiagent Systems Intelligence Based on Paired Intelligence Measurements, International Journal of Intelligent Systems, 33(3), 2018, pp. 463-486

[4] Iantovics, L. B. Black-Box-Based Mathematical Modelling of Machine Intelligence Measuring, Mathematics, 9(6), 2021, 681

[5] Arik, S., Iantovics, L. B., Szilagyi, S. M. OutIntSys - a Novel Method for the Detection of the Most Intelligent Cooperative Multiagent Systems, 24th Int. Conf. on Neural Information Processing (ICONIP 2017), 14-18 Nov. 2017, Guangzhou, China. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (Eds.), Neural Information Processing, Lecture Notes in Computer Science 10637, 2017, pp. 31-40

[6] Chen, S. H., Zhou, X. Q., Zhou, G., Fan, C. L., Ding, P. X., Chen, Q. L. An online physical-based multiple linear regression model for building’s hourly cooling load prediction. Energy and buildings, 254, 2022, 111574

[7] Leij, F. J., Dane, J. H., Sciortino, A. Hierarchical prediction of soil water content time series. Catena, 209(2), 2022, 105841

[8] Soltani, E., Rezaei, M., Nasiri, M., Barasteh, S., Rahmati-Najarkolaei, F., Mazaheri, MA. The Bivariate Correlation of Health Literacy and Cell Phone Addiction amongst Iranian Healthcare Students, Journal of Clinical and Diagnostic Research, 13(6), 2019, IC1-IC5

[9] Indrayan, A. Statistical fallacies & errors can also jeopardize life & health of many, Indian J Med Res. 148(6), 2018, 677-679

[10] Zeng, Y. X., Dong, J. J., Ji, Z. J., Yang, C. D., Liang, Y. Linear Regression Model for the Prediction of Rice Sheath Blight Field Resistance. Plant Disease, 105(10), 2021, pp. 2964-2969

[11] Biswas, N. K., Banerjee, S., Biswas, U., Ghosh, U. An approach towards development of new linear regression prediction model for reduced energy consumption and SLA violation in the domain of green cloud computing. Sustainable Energy Technologies and Assessments, 45, 2021, 101087

[12] Machova, K., Mach, M., Hreskova, M., Classification of Special Web Reviewers Based on Various Regression Methods, Acta Polytechnica Hungarica, 17(3), 2020, pp. 229-248

[13] Atobishi, T., Bahna, M., Takacs-Gyorgy, K., Fogarassy, C., Factors Affecting the Decision of Adoption Cloud Computing Technology: The Case of Jordanian Business Organizations, Acta Polytechnica Hungarica, 18(5), 2021, pp. 131-154

[14] Jadlovska, A., Jajcisin, S., Predictive Control Algorithms Verification on the Laboratory Helicopter Model, Acta Polytechnica Hungarica, 9(4), 2012, pp. 221-245

[15] Artin, J., Valizadeh, A., Ahmadi, M., Kumar, S. A. P., Sharifi, A. Presentation of a Novel Method for Prediction of Traffic with Climate Condition Based on Ensemble Learning of Neural Architecture Search (NAS) and Linear Regression. Complexity, 2021, 8500572

[16] Kim, T., Sharda, S., Zhou, X. S., Pendyala, R. M., A stepwise interpretable machine learning framework using linear regression (LR) and long short-term memory (LSTM): City-wide demand-side prediction of yellow taxi and for-hire vehicle (FHV) service. Transportation Research Part C: Emerging Technologies, 120, 2020, 102786

[17] Lei, X. J., Zhang, W. X. Logistic regression algorithm to identify candidate disease genes based on reliable protein-protein interaction network. Science China Information Sciences, 64(7), 2021, 179101

[18] Liang, J. Y., Song, Y. S., Li, D. Y., Wang, Z. Q., Dang, C. Y., An accelerator for the logistic regression algorithm based on sampling on-demand. Science China Information Sciences, 63(6), 2020, 169102

[19] Wei, D. P., Wang, T., Wang, J. A logistic regression model for Semantic Web service matchmaking. Science China Information Sciences, 55(7), 2012, pp. 1715-1720

[20] Iantovics, L. B., Enăchescu, C., Method for Data Quality Assessment of Synthetic Industrial Data, Sensors 22(4), 2022, 1608

[21] Madadizadeh, F., Ezati Asar, M., Hosseini, M., Common Statistical Mistakes in Descriptive Statistics Reports of Normal and Non-Normal Variables in Biomedical Sciences Research, Iranian Journal of Public Health, 44 (11), 2015, pp. 1557-1558

[22] Iantovics, L. B., Dehmer, M., Emmert-Streib, F. MetrIntSimil-An Accurate and Robust Metric for Comparison of Similarity in Intelligence of Any Number of Cooperative Multiagent Systems, Symmetry, 10(2), 2018, 48

[23] King, G. How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science, American Journal of Political Science, 30(3), 1986, pp.666-687

[24] Marusteri, M., Bacarea, V. Comparing groups for statistical differences: How to choose the right statistical test?, Biochemia Medica, 20(1), 2010, 15-32

[25] Karadeniz, P. G., Uzabacı, E., Kuyuk, S. A., Kesin, F. K., Can, F. E., Seçil, M., Ercan, İ. Statistical errors in articles published in radiology journals. Diagn Interv Radiol 25(2), 2019, 102-108

[26] Raheem, Y. A. Statistics in medical research: Common mistakes, J Taibah Univ Med Sci. 18(6), 2023, 1197-1199

[27] Rovetta A. Common Statistical Errors in Scientific Investigations: A Simple Guide to Avoid Unfounded Decisions. Cureus. 15(1),2023, e33351

[28] Ignatiadis, N., Saha, S., Sun, D. L., Muralidharan, O., Empirical Bayes Mean Estimation With Nonparametric Errors Via Order Statistic Regression on Replicated Data, Journal of the American Statistical Association, 118(542), 2021, pp. 987-999

[29] Ranganai, E., Nadarajah, S. A predictive leverage statistic for quantile regression with measurement errors, Communications in Statistics - Simulation and Computation, 46(8), 2017, pp. 6385-6398

[30] Abuzaid, A. H., Hussin, A. G., Mohamed, I. B., Detection of outliers in simple circular regression models using the mean circular error statistic, Journal of Statistical Computation and Simulation, 83(2), 2013, pp. 269-277

[31] Chang, X. F., Yang, H. Performance of the preliminary test two-parameter estimators based on the conflicting test statistics in a regression model with Student's t error, Statistics, 46(3), 2012, pp. 291-303

[32] Serao, N. V., Tokach, M. D., Paton, N. Fundamentals, Common Mistakes, and Graduate Education in Statistics, Journal of Animal Science, 99, 2021, pp. 104-104

[33] Holmes, J. B., Speed, D., Balding, D. J., Summary statistic analyses can mistake confounding bias for heritability, Genetic epidemiology, 43(8), 2019, pp. 930-940

[34] Kalantar, A., Gelb, R. I., Alper, J. S. Biases in summary statistics of slopes and intercepts in linear regression with errors in both variables. Talanta, 42(4), 1995, pp. 597-603

[35] Hulley S. B., Cummings S. R., Browner W. S., Grady D., Newman T. B. Designing clinical research: an epidemiologic approach. 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2013. Appendix 6C

[36] Shapiro, S. S., Wilk, M. B. An analysis of variance test for normality (complete samples). Biometrika, 52, 1965, pp. 591-611

[37] Razali, N., Wah, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modelling and Analytics, 2, 2011, pp. 21-33

[38] Dallal, G. E., Wilkinson, L. An analytic approximation to the distribution of Lilliefors's test statistic for normality. American Statistician, 40, 1986, pp. 294-296

[39] Lilliefors, H. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. 62, 1967, pp. 399-402

[40] Lilliefors, H. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. J. Am. Stat. Assoc. 64, 1969, pp. 387-389

[41] Chakravarti, I. M., Laha, R. G., Roy, J. Handbook of Methods of Applied Statistics; Wiley: New York, NY, USA, 1967, Vol. I, pp. 392-394

[42] Tsai, D. M., Yang, C. H., A quantile-quantile plot based pattern matching for defect detection, Pattern Recognition Letters, 26(13), 2005, pp.1948-1962

[43] Ben, M. G., Yohai, V. J., Quantile-quantile plot for deviance residuals in the generalized linear model, Pattern Recognition Letters, 13(1), 2004, pp. 36-47

[44] Bonett, D. G., Wright, T. A. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. Psychometrika, 65, 2000, pp. 23-28

[45] Stigler, S. M. Francis Galton's Account of the Invention of Correlation. Statistical Science, 4(2), 1989, pp. 73-79

[46] Chicco, D., Warrens, M. J., Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7(e623), 2021, e623

[47] Meloun M., Hill, M., Militký, J., Vrbíková, J, Stanická, S, Skrha, J. New methodology of influential point detection in regression model building for the prediction of metabolic clearance rate of glucose. Clin Chem Lab Med. 42(3), 2004, 311-322

[48] Sauerbrei, W., Buchholz, A., Boulesteix, A. L., Binder, H. On stability issues in deriving multivariable regression models. Biom J. 57(4), 2015, 531-555