Empirically Supported Methodological Critique of Double Entry in Dyadic Data Analysis

Imre Dobos[1] | *Budapest University of Technology and Economics, Budapest, Hungary*
Andrea Gelei[2] | *Corvinus University of Budapest, Budapest, Hungary*

Abstract

Analyzing dyadic phenomena (e.g. trust, power, and satisfaction) gains importance not only in sociology and psychology, but also in economics and management. The aim of the paper is to examine the mathematical foundation of Dyadic Data Analysis (DDA). On one hand, we critique the database development of DDA for exchangeable cases, and develop an algorithm for transforming such a data set into distinguishable cases. On the other hand, we question the usefulness of a widely used data development technique of DDA, the so-called double entry. We reason that this technique does not necessarily lead to additional information. In contrast, it might lead to information losses. We develop approximations for correlations and regression models of DDA. These are also empirically tested using a database of 89 dyads. The obtained results back our theoretical reasoning, most of the approximations give satisfying results. This support our main proposition that mathematical foundation of DDA needs further research.

Keywords

Correlational analysis, regression analysis, dyadic data analysis, double entry technique

JEL code

C10, C39, C49

INTRODUCTION

The problem of analyzing dyadic data is well known from paired samples. The basic question is whether or not a given variable in two dependent samples has the same shape of distribution, expected value, and standard deviation. These questions are important but are also supplemented by new research challenges, since researchers in sociology, psychology, economics and management are increasingly interested in complex research issues that make it necessary to apply multivariate analytical techniques in dyadic settings (Kenny et al., 2006). The traditional technique of paired sample analysis is inappropriate for answering such research questions. (e.g. How the level of perceived trust of the partners in a business relationship influences the partners' willingness to take risk in joint future innovation projects.) Instead,

[1] Budapest University of Technology and Economics, Magyar Tudósok Körútja 2, 1111 Budapest, Hungary. Corresponding author: email: dobos@kgt.bme.hu.
[2] Corvinus University of Economics, Fővám tér 8, 1093 Budapest, Hungary.

the use of dyadic data analysis (DDA) is suggested (Griffin and Gonzalez, 1995; Gonzalez and Griffin, 1999; 2000; Gonzales, 2010; Kenny et al., 2006; Burk et al., 2007; Kenny, 2015). According to literature, the processing of paired samples, also called dyadic databases, using traditional statistical methods may lead to a number of error types[3] (Gonzalez and Griffin, 2000). Dyadic data analysis is a specific, interrelated set of statistical techniques that aim at overcoming these errors.

Moreover, let us point out limitation of traditional inductive statistics, namely it assumes the representativeness of the sample. In an explicit form, this usually appears in a way that data observed can be regarded as the independent sample with identical distribution, or it can be assumed that it approaches identical distribution because of the small sample and different placements of weight. In the examination of relational trust and similar social problems, the representativeness of the sample is out of the question. Often, the population is not known by the researcher, the respondents are the ones who just participate in the given study, which means that the analysis is basically descriptive. In this case, inductive statistics makes little sense. The essence of the dyadic approach is that it regards each relationship unique and intends to put the consequences of the unique context in the center of analysis. Hence, this approach does not pose any requirements about generalization regarding total population either (Gelei and Sugár, 2017).

Previously dyadic data analysis to a trust-related management problem has already been applied (Gelei and Dobos, 2016). Later, DDA and the classical statistical techniques have been compared (Gelei and Sugár, 2017). This comparison concluded, despite the substantial methodological differences between classical statistics and DDA, that the empirical results were not significantly different. This finding has motivated us to look into the mathematical fundamentals of dyadic data analysis. The results of this investigation make up this methodological article. We discuss key concepts of dyadic data analysis, focusing on the suggested database development technique, called double entry (Gonzalez and Griffin, 2000; Ledermann and Kenny, 2015). We discuss the so-called exchangeable case, the related homogeneity analysis, the core correlations of DDA and its regression equations (Gonzalez and Griffin, 1995, 1999; Ledermann et al., 2011). These fundamentals are relevant for more elaborate and complex analytical techniques, such as the curve-of-factors model (McArdle, 1988; Whittaker et al., 2014), structural equation modeling (Peugh et al., 2013; Deng and Yan, 2015), and situations dealing with longitudinal dyadic data (Planalp et al., 2017). Our objective is to critically discuss this relatively new statistical methodology.

In dyadic data analysis, the first analytical step is the so-called homogeneity analysis. Here, one deals with the problem of assessing interdependence in a dyad for a single variable (Gonzalez and Griffin, 2000). The key question is whether the informants in a dyad have symmetric or asymmetric positions (e.g., physician and patient). First, we talk about exchangeable cases, in contrast to distinguishable ones. The homogeneity analysis is different from the classical analysis, in which the core issue is to evaluate the similarity of the distributions of the variables in two databases. Instead of using the ANOVA framework, DDA suggests applying a technique for database development called double entry (Gonzalez and Griffin, 2000). This technique has crucial importance not only for DDA but also for our critique. Therefore, in the following sections, we discuss basic concepts and techniques of dyadic data analysis, including double entry and homogeneity analysis. As a next step, we attempt to refine the concept of dyadic correlations and calculate them using the initial, raw database, which does not necessitate the use of the double-entry technique. (This database reflects the timely development of pairwise sampling; the first pair in the survey is fixed in the database as the first dyad, and so on.) Finally, dyadic regression models are investigated. We conclude that the suggested technique of double entry and the statistical constructs using these models do not necessarily lead to additional information. In contrast, these techniques might lead to information

[3] These error types are the following: (1) error of assumed independence; (2) data omission error; (3) error between levels; and (4) error of the levels of analysis (Gonzalez and Griffin, 2000).

losses. We develop statistical approximations for these dyadic constructs by applying classical statistics on the initial, raw database.

Each of the suggested statistical constructs are tested using a database that has been developed in one of our previous field studies using pairwise sampling. The research hypothesis of this previous study was as follows: In a business relationship characterized by mutually high levels of trustworthiness perceived by the counterparts, the willingness to be involved in risky situations is higher than in relationships in which actors do not mutually believe that their partners are highly trustworthy. In order to test our hypothesis, we developed a questionnaire, where respondents had to answer the followings:

- Evaluate the perceived levels of trustworthiness of their actual pair (1–7 Likert scale);
- Evaluate the level of different information sharing situations listed (ranking);
- Trust in the relationship: the willingness to share specific information with the actual partner in the pair (yes = 1 or no = 0).

We organized workshops for purchasing and logistics managers, where theyir formed concrete pairs and filled out the questionnaire. Data gathering was so carried out in the physical presence of respondents, but in an anonym way. Concrete answers were neither visible nor accessible to the participants in order to avoid biases in responses. We gathered 89 pairs of questionnaires, with 178 dyadic data points. A more detailed description of the field study and its dataset development is presented in the work of Gelei and Dobos (2016). For this article we used the variable of perceived levels of trustworthiness for calculating the newly developed correlation constructions based on the initial, raw dataset. For testing our suggested regression models we used trust as the dependent variable while independent variables were the perceived levels of trustworthiness in the pairs. We used SPSS 22 and Microsoft Excel throughout this article for statistical calculations.

The results of our empirical test show that in most cases, the suggested approximations can give really good results and support our suggestion not to use the difficult technique of double entry and the statistical constructs based on it.

1 FUNDAMENTALS OF THE CRITIQUE OF DYADIC DATA ANALYSIS

These form the analytical unit for statistical analysis. A very simple question arises, when developing dyadic datasets from such data pairs: In what order to fix the two answers of a pair? In a distinguishable case it is obvious since positions in any pair are given (e.g. doctor and patient). An initial or raw database is shown in Table 1.

Table 1 Dyadic data analysis with three dyads in the database

Variables **Observations**	***1. variable (X)***	
	1. data (X_1)	***2. data (X_2)***
1. dyad	x_{11}	x_{12}
2. dyad	x_{21}	x_{22}
3. dyad	x_{31}	x_{32}

Source: Own construction

In the so-called exchangeable case however these positions are not predefined, and can change. In such cases n number of such data pairs can lead to a number of 2^n number databases. In case data pairs are interpreted as paired sample, different datasets can lead to different results during analysis. Therefore,

the question arises, which one should be use? This is the first problem we investigate. We suggest a method which transforms any exchangeable data set into a distinguishable one. As a next step another issue related to dyadic dataset development is discussed, the so-called double entry. Our focal problem is, whether this doubling leads to any information surplus or not.

1.1 Issues of data set development in DDA

A key innovation in dyadic data analysis is the double entry of data obtained through field research using pairwise sampling. The essential idea is to create two vectors from all the aligned data pairs by changing the order in which the data are entered into the database. Changing this order creates two variables from one. The original and the newly created variables are denoted as X and X'; see the example in Table 2. This table shows that the number of observations belonging to variables X and X' is twice the number of dyads, which is the number of pairs in the database. Dyadic data analysis requires this transformation to create and use vectors instead of matrices (tables) for further statistical analysis.

Table 2 Symbolic representation for double entry and the pairwise data setup

Observations	***Variables***	
	X	***X'***
1. pair (initial order)	x_{11}	x_{12}
1. pair (changed order)	x_{12}	x_{11}
2. pair (initial order)	x_{21}	x_{22}
2. pair (changed order)	x_{22}	x_{21}
3. pair (initial order)	x_{31}	x_{32}
3. pair (changed order)	x_{32}	x_{31}
4. pair (initial order)	x_{41}	x_{42}
4. pair (changed order)	x_{42}	x_{41}

Source: Own construction

1.2 The so-called exchangeable case and homogeneity analysis

In DDA, there are two types of analytical situations called cases, including the exchangeable and the distinguishable cases (Gonzalez and Griffin, 2000). In the exchangeable case, the informants in a given dyad (or pair) cannot be distinguished in advance, in contrast to the distinguishable case, in which the informants in any given dyad have specific systemic characteristics or positions that are known well in advance of the analysis (e.g., one informant in a pair is the husband, the other is the wife). In this article, the analysis starts with the exchangeable case.

As mentioned previously, in the distinguishable case, the two people in a pair are in asymmetric positions, in contrast to the exchangeable case, in which the positions of the two informants in any pair are symmetric, i.e., they are identical. Suppose we have three dyads or pairs in the database, as shown in Table 1. This table reflects the sequential data collection in field research: the first dyad (or pair) was the first one questioned, the second dyad was questioned next, and so on.

Since we have exchangeable cases, we can transform this initial database by simply changing the order of the data related to a given variable in a dyad.

Table 3 Changing the order of the data related to a given variable in a dyad to develop a new database for dyadic data analysis

Variables **Observations**	***1. variable (X)***	
	1. data (X'$_1$)	***2. data (X'$_2$)***
1. dyad	x_{12}	x_{11}
2. dyad	x_{21}	x_{22}
3. dyad	x_{31}	x_{32}

Source: Own construction

Since we have three dyads, this process could be continued an additional six time periods, leading to $2^3 = 8$ potential databases. Generally, we can state that having n dyads for analysis offers 2^n slightly different databases.

Table 4 Potential databases with three dyads in the survey

	Database 1	**Database 2**	**Database 3**	**Database 4**	**Database 5**	**Database 6**	**Database 7**	**Database 8**
1. dyad	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{12}, x_{11})	(x_{12}, x_{11})	(x_{12}, x_{11})	(x_{12}, x_{11})
2. dyad	(x_{21}, x_{22})	(x_{21}, x_{22})	(x_{22}, x_{21})	(x_{22}, x_{21})	(x_{21}, x_{22})	(x_{21}, x_{22})	(x_{22}, x_{21})	(x_{22}, x_{21})
3. dyad	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})

Source: Own construction

To obtain valid and reliable results, the statistical analysis applied to dyadic data must be unaffected by the choice of database. Let us test this first! For this purpose, we used the database with 89 purchasing and logistics manager dyads.

We chose the first dataset for our calculation randomly from all the potential databases, while the second was developed from this initial one by carrying out the following change systematically: the data with lower value in any given dyad/pair were recorded systematically in data position 1 in the dyad. Since we can assume that the two datasets are interdependent, we applied a test developed for paired samples, namely, the *t*-test. Table 5 shows that the results obtained using the two databases are significantly different. The first database led to the acceptance of the null hypothesis; the means of the two informants of the pairs in the given database do not differ significantly. In contrast, using the second, modified database resulted in the rejection of this null hypothesis. The objective was to highlight the problem related to the choice of database which might lead markedly different results.

So, the order of the data in the databases might really pose problems in exchangeable cases. According to dyadic data analysis, a potential solution to this challenge could be the technique of double entry. However, this solution does not really solve the problem; it only doubles the size of the database. As discussed above, the number of potential databases is 2^n, since we have n dyads available for analysis. Any statistical method applied for analyzing dyadic data must be completely independent from the order

Table 5 Testing the means of the two databases

	Paired differences					t-test	Freedom	Level of significance (two-tailed)
	Mean	Standard deviation	Standard error	Confidence interval (95%) for the differences				
				Lower	Upper			
Database 1	0.07865	1.79788	0.19058	–0.30008	0.45738	0.413	88	0.681
Database 2	1.13483	1.39146	0.14749	0.84172	1.42795	7.694	88	0.000

Source: Own construction

of the data in the database, i.e. the database is transformed into a distinguishable one. Let us note that such methods can rely on a technique that operates on the absolute values of the sum and/or difference of the data in a given dyad; these are the same for all dyads.

Considering this let us introduce two new variables, z_{i1} and z_{i2} from the raw database, as follows:

$$z_{i1} = ½ (x_{i1} + x_{i2}), \text{ and } z_{i2} = ½ |x_{i1} - x_{i2}|,$$

where x_{i1} and x_{i2} are the answers of the first and second informants of the dyads to a specific question. The first new variable (z_{i1}) can be interpreted as the variable measuring the aggregated effect, while the second (z_{i2}) measures the differences between these answers. We emphasize that the benefit of these new variables subsists in their indifference to the order in which the data are entered into the database. One can easily recover the initial, raw data from these new variables:

a) If $x_{i1} \geq x_{i2}$, then $x_{i1} = z_{i1} + z_{i2}$ and $x_{i2} = z_{i1} - z_{i2}$.
b) If $x_{i1} < x_{i2}$, then $x_{i2} = z_{i1} + z_{i2}$ and $x_{i1} = z_{i1} - z_{i2}$.

1.3 Homogeneity analysis of exchangeable cases – the pairwise intraclass correlation

In exchangeable cases, one should begin dyadic data analysis with homogeneity analysis, which is carried out using the pairwise interclass correlation (Kenny et al., 2006) based on double entry. As stated above, the null hypothesis is that the informants of the dyads give homogeneous answers. In this section, we demonstrate that homogeneity can be tested in a simpler way, based on the initial database that does not require the technique of double entry. We develop and suggest a formula that approximates the suggested pairwise interclass correlation of DDA. After presenting our theoretical argument, we test the developed formula using our database and calculate the homogeneity in the case of randomly chosen variables with both the pairwise interclass correlation and our suggested approximation.

1.3.1 Theoretical argument

As described above, the technique of double entry transforms the initial database with n dyads (vectors) into another database with $2n$ dyads, as in Table 2. Suppose we fix the order of the informants within the dyads. This means that the previously mentioned issue of exchangeability is not a problem. We denote the values of the variables in the initial database as (x_1, x_2) and (y_1, y_2). The values of the same variables obtained with double entry are denoted as (X, X') and (Y, Y'). The values (X, X') and (Y, Y') are the transformed values of (x_1, x_2) and (y_1, y_2). Therefore, assuming the data can be rearranged, we obtain:

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, X' = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \text{ and } Y' = \begin{bmatrix} y_2 \\ y_1 \end{bmatrix}.$$

This equation reflects that the new variables can be derived from the initial ones by arranging the two vectors for a specific observation, one on top of the other, in reverse order. Remember, our key question is whether applying double entry is beneficial or not; do we obtain additional information with this method that is useful for further statistical analysis?

To avoid biases, we assume that the vectors represent the population. This way, we can facilitate calculation and use the number of vectors in variance-covariance calculations. First, we calculate the means for both variables and databases:

$$E(X) = E(X') = \frac{E(x_1) + E(x_2)}{2}, \text{ and } E(Y) = E(Y') = \frac{E(y_1) + E(y_2)}{2},$$

which can be determined easily. These equations indicate that the means of the new variables obtained through double entry are equal to the means of the original elements. We can formulate this differently; the mean of all the answers corresponding to a variable is the same as the mean of vectors X and Y, which stems from the technique of double entry.

Calculating the variance requires slightly more patience, but it is not very complicated either:

$$var(X) = var(X') = \frac{var(x_1) + var(x_2)}{2} + \left(\frac{E(x_1) + E(x_2)}{2}\right)^2, \text{ and} \tag{1}$$

$$var(Y) = var(Y') = \frac{var(y_1) + var(y_2)}{2} + \left(\frac{E(y_1) + E(y_2)}{2}\right)^2. \tag{2}$$

In addition, the covariance is calculated as follows:

$$cov(X, X') = cov(x_1, x_2) - \left(\frac{E(x_1) + E(x_2)}{2}\right)^2, \text{ and} \tag{3}$$

$$cov(Y, Y') = cov(y_1, y_2) - \left(\frac{E(y_1) + E(y_2)}{2}\right)^2. \tag{4}$$

Moreover,

$$cov(X, Y) = cov(X', Y') = \frac{cov(x_1, y_1) + cov(x_2, y_2)}{2} + \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}, \text{ and} \tag{5}$$

$$cov(X, Y') = cov(X', Y) = \frac{cov(x_1, y_2) + cov(x_2, y_1)}{2} - \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}. \tag{6}$$

Let us note here that the double entry actually decreases the amount of useful information, since the mean, variance and covariance of the new variables are, in several cases, the same. Because of the symmetries mentioned, related indices of variables (x_1, x_2) and (y_1, y_2) cannot be calculated from the new variables (X, X') and (Y, Y') without knowing a construction algorithm of the last variables. This finding reflects a unidirectional logical relationship between the two databases; variables (x_1, x_2) and (y_1, y_2) unambiguously determine (X, X') and (Y, Y'), while the reverse does not hold. The loss of information is due to this asymmetry.

This fact has the consequence that we can find a relation between the new and old covariances only in a few cases. These cases are the following using (3) and (4):

$$cov(X, X') \leq cov(x_1, x_2) \text{ and } cov(Y, Y') \leq cov(y_1, y_2).$$

Additionally, the variance is a special case of the covariance from (1) and (2):

$$var\,(X) \geq \frac{var\,(x_1) + var\,(x_2)}{2} \geq \sqrt{var\,(x_1)} \cdot \sqrt{var\,(x_2)}\,, \text{ and}$$

$$var\,(Y) \geq \frac{var\,(y_1) + var\,(y_2)}{2} \geq \sqrt{var\,(y_1)} \cdot \sqrt{var\,(y_2)}\,.$$

We suppose that the informants in any dyad give nearly the same answers, i.e. the expected values are nearly the same. This can be expressed as follows:

$$max\,\{|E\,(x_1) - E\,(x_2)|;\, |E\,(y_1) - E\,(y_2)|\} \leq \varepsilon\,,$$

where ε is an arbitrarily small positive number. In this way, using the initial data, we obtain the following approximations for the new variables, which we obtained by using double entry:

$$var\,(X) = var\,(X') \sim \frac{var\,(x_1) + var\,(x_2)}{2},$$

$$var\,(Y) = var\,(Y') \sim \frac{var\,(y_1) + var\,(y_2)}{2},$$

$$cov\,(X,\, X') \sim cov\,(x_1, x_2)\,,$$

$$cov\,(Y,\, Y') \sim cov\,(y_1, y_2)\,,$$

$$cov\,(X,\, Y) = cov\,(X',\, Y') \sim \frac{cov\,(x_1, y_1) + cov\,(x_2, y_2)}{2}, \text{ and}$$

$$cov\,(X,\, Y') = cov\,(X',\, Y) \sim \frac{cov\,(x_1, y_2) + cov\,(x_2, y_1)}{2}.$$

These relations can be confirmed using elementary statistical methods, so we do not present their detailed derivation. We can state that the variance of variable X is larger than the product of the variances of the two vectors. This also might lead to a loss of information.

Since in case of the two covariances – $cov\,(X,Y)$ and cov (X,Y') – the product of the expected values on the right-hand side can be either positive or negative, we cannot estimate the relation between the covariances. However, we can state that:

$$cov\,(X,\, Y) + cov\,(X,\, Y') = cov\,(X,\, Y + Y') = \frac{cov\,(x_1, y_1) + cov\,(x_1, y_2) + cov\,(x_2, y_1) + cov\,(x_2, y_2)}{2} = \frac{cov\,(x_1 + x_2, y_1 + y_2)}{2},$$

which results from the application of variance-covariance algebra.

We express the two correlations using the following formulas using (3), (1) and (4), (2):

$$r\,(X, X') = \frac{\sqrt{var\,(x_1)} \cdot \sqrt{var\,(x_2)} \cdot r\,(x_1, x_2) - \left(\frac{E\,(x_1) - E\,(x_2)}{2}\right)^2}{\frac{var\,(x_1) + var\,(x_2)}{2} + \left(\frac{E\,(x_1) - E\,(x_2)}{2}\right)^2}, \text{ and}$$

$$r\,(Y,\,Y') = \frac{\sqrt{var\,(y_1)} \cdot \sqrt{var\,(y_2)} \cdot r\,(y_1, y_2) - \left(\frac{E\,(y_1) - E\,(y_2)}{2}\right)^2}{\frac{var\,(y_1) + var\,(y_2)}{2} + \left(\frac{E\,(y_1) - E\,(y_2)}{2}\right)^2}.$$

Recall that the variances of the two new variable pairs (X' and Y') are the same as variances of X and Y. Therefore, the correlations can be approximated as follows in case of positive correlations:

$$r\,(X,\,X') = \frac{\sqrt{var\,(x_1)} \cdot \sqrt{var\,(x_2)}}{\frac{var\,(x_1) + var\,(x_2)}{2}} \cdot r\,(x_1, x_2) \leq r\,(x_1, x_2), \text{ and} \tag{7}$$

$$r\,(Y,\,Y') = \frac{\sqrt{var\,(y_1)} \cdot \sqrt{var\,(y_2)}}{\frac{var\,(y_1) + var\,(y_2)}{2}} \cdot r\,(y_1, y_2) \leq r\,(y_1, y_2).$$

This equation implies that the homogeneity analysis of dyadic data analysis can be carried out not only using the ANOVA tables but also using the initial database. There is no need to introduce new variables by applying double entry. Using correlations $r\,(x_1, x_2)$ and $r\,(y_1, y_2)$,we can analyze whether the answers of the two informants in a given dyad correspond to each other or not, i.e., whether a linear relationship between them exists or not. The suggested method and the same calculations are also relevant for distinguishable cases.

1.3.2 Testing homogeneity with DDA and the suggested approximation

Above, we presented a new formula that can replace pairwise intraclass correlation so double entry can be omitted. In this way, statistical analysis becomes easier yet remains reliable. Recall that this formula is given as (7).

This formula not only indicates the possibility of leaving out double entry but also reveals that it will result in higher values than the original dyadic correlation based on double entry. This finding may also indicate information loss due to double entry.

We tested the homogeneity using both formulas. The original pairwise intraclass correlation index is 0.490537. The reduced, simplified correlation index is 0.490877. This supports our statement that the difficulties raised by double entry are not outweighed by its potential positive effect.

2 A CRITICAL DISCUSSION OF CORRELATIONS OF DYADIC DATA ANALYSIS

Dyadic data analysis has introduced five types of correlations (Griffin and Gonzalez, 1995, 1999, 2004), excluding the pairwise interclass correlation discussed above:

1. Overall within-partner correlation;
2. Cross-intraclass correlation;
3. Mean-level correlation (correlation between dyad means);
4. Individual-level correlation;
5. Dyad-level correlation.

This section critically discusses these correlations and presents approximations for them based on a similar logic to that applied before. First, we theoretically discuss these correlations, and develop the approximations. Next, we test them using the database developed previously.

2.1 The overall within-partner and the cross-intraclass correlations

The overall within-partner correlation $r(X, Y)$ is specified in dyadic data analysis by the following equation using (5), (1) and (2):

$$r(X, Y) = \frac{\frac{1}{2} \cdot [\sqrt{var(x_1)} \cdot \sqrt{var(y_1)} \cdot r(x_1, y_1) + \sqrt{var(x_2)} \cdot \sqrt{var(y_2)} \cdot r(x_2, y_2)] + \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}}{\sqrt{\frac{var(x_1) + var(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \cdot \sqrt{\frac{var(y_1) + var(y_2)}{2} + \left(\frac{E(y_1) - E(y_2)}{2}\right)^2}}.$$

The covariance in the formula's numerator measures the direction of the stochastic relationship between the answers of the two informants within a given dyad. In this way, this covariance can be interpreted as an 'internal' or 'individual' correlation.

Let us suppose again that both expected values and variances are approximately the same. Then,

$$max\{|var(x_1) - var(x_2)|; |var(y_1) - var(y_2)|\} \leq \eta,$$

where η is an arbitrarily small number. For the above positive correlation, we can formulate the following approximation:

$$r(X, Y) \sim \frac{\frac{1}{2} \cdot [\sqrt{var(x_1)} \cdot \sqrt{var(y_1)} \cdot r(x_1, y_1) + \sqrt{var(x_2)} \cdot \sqrt{var(y_2)} \cdot r(x_2, y_2)]}{\sqrt{\frac{var(x_1) + var(x_2)}{2}} \cdot \sqrt{\frac{var(y_1) + var(y_2)}{2}}}$$
$$\leq \frac{1}{2} \cdot [r(x_1, y_1) + r(x_2, y_2)].$$

The cross-intraclass correlation is defined as follows using (6), (1) and (2):

$$r(X, Y') = \frac{\frac{1}{2} \cdot [\sqrt{var(x_1)} \cdot \sqrt{var(y_2)} \cdot r(x_1, y_2) + \sqrt{var(x_2)} \cdot \sqrt{var(y_1)} \cdot r(x_2, y_1)] + \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}}{\sqrt{\frac{var(x_1) + var(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \cdot \sqrt{\frac{var(y_1) + var(y_2)}{2} + \left(\frac{E(y_1) - E(y_2)}{2}\right)^2}}.$$

The covariance of the initial dataset reflects the relationship between the answers given by the two informants of a specific dyad to two different questions. Based on the previous argument, this covariance is approximated as follows:

$$r(X, Y') \sim \frac{\frac{1}{2} \cdot [\sqrt{var(x_1)} \cdot \sqrt{var(y_2)} \cdot r(x_1, y_2) + \sqrt{var(x_2)} \cdot \sqrt{var(y_1)} \cdot r(x_2, y_1)]}{\sqrt{\frac{var(x_1) + var(x_2)}{2}} \cdot \sqrt{\frac{var(y_1) + var(y_2)}{2}}}$$
$$\leq \frac{1}{2} \cdot [r(x_1, y_2) + r(x_2, y_1)].$$

2.2 Mean-level correlation

The mean-level correlation (also called the correlation between dyad means) is specified by Griffin and Gonzalez (1995) as follows:

$$r_m(X, X', Y, Y') = \frac{r(X, Y) + r(X, Y')}{\sqrt{1 + r(X, X')} \cdot \sqrt{1 + r(Y, Y')}}. \quad (9)$$

The Formula (9) can be rewritten in terms of variances and covariances. After small transformations, and using elementary covariance algebra, we obtain:

$$r_m(X, X', Y, Y') = \frac{cov(X, Y + Y')}{\sqrt{cov(X, X + X')} \cdot \sqrt{cov(Y, Y + Y')}}.$$

After calculating the covariance, this expression can be rewritten in terms of the raw data:

$$r_m(X, X', Y, Y') = \frac{\frac{1}{2} \cdot cov(x_1 + x_2, y_1, y_2)}{\sqrt{\frac{1}{2} \cdot var(x_1 + x_2)} \cdot \sqrt{\frac{1}{2} \cdot var(y_1 + y_2)}} = r(x_1 + x_2, y_1, y_2).$$

This means that the dyad-level correlation is a classical correlation that interprets the correlation between two newly introduced variables as the sum of the dyad-level values. Interestingly, when using new data, $r_m(X, X', Y, Y')$ does not correspond to the traditional Pearson correlation because the covariance in the numerator suggests the formula $\sqrt{var(X)} \cdot \sqrt{var(Y + Y')}$ instead of the covariance. If somebody takes the trouble to calculate the classical correlation, he/she will conclude:

$$r(X, Y + Y') = \frac{cov(X, Y + Y')}{\sqrt{var(X)} \cdot \sqrt{var(Y + Y')}} = \frac{\frac{1}{2} \cdot cov(x_1 + x_2, y_1, y_2)}{\sqrt{\frac{var(x_1) + var(x_2)}{2} + \sqrt{\left(\frac{E(x_1) - E(x_2)}{2}\right)^2}} \cdot \sqrt{\frac{1}{2} \cdot var(y_1 + y_2)}}.$$

This is not the same as the previous correlation, $r(x_1 + x_2, y_1 + y_2)$, but it is very close to it.

2.3 Individual-level correlation

The most problematic correlation coefficients in dyadic data analysis are the individual- and dyad-level correlation coefficients. The individual-level correlation is suggested to calculate:

$$r_i(X, X', Y, Y') = \frac{r(X, Y) - r(X, Y')}{\sqrt{1 - r(X, X')} \cdot \sqrt{1 - r(Y, Y')}}. \quad (10)$$

The Formula (10) can also be rewritten in terms of the variance and covariance:

$$r_i(X, X', Y, Y') = \frac{cov(X, Y - Y')}{\sqrt{cov(X, X - X')} \cdot \sqrt{cov(Y, Y - Y')}}.$$

Before proceeding with the transformation, we present the traditional Pearson correlation, which is widely available in the statistical literature:

$$r(X, Y - Y') = \frac{cov(X, Y - Y')}{\sqrt{var(X)} \cdot \sqrt{var(Y - Y')}} = \frac{\frac{1}{2} \cdot cov(x_1 - x_2, y_1 - y_2) + \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}}{\sqrt{\frac{var(x_1) + var(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \cdot \sqrt{var(y_1 - y_2) + E(y_1) - E(y_2))^2}}.$$

Now, we continue the process of reducing the correlation to a formula using initial, raw data. The above expression is similar to the mean-level correlation discussed above; the difference is in the reversed signs. As a next step, we substitute our initial data into the above formula and

$$r_i(X, X', Y, Y') = \frac{\sqrt{var(x_1 - x_2)} \cdot \sqrt{var(y_1 - y_2)} \cdot r(x_1 - x_2, y_1 - y_2) + [E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{\sqrt{var(x_1 - x_2) + [E(x_1) - E(x_2)]^2} \cdot \sqrt{var(y_1 - y_2) + [E(y_1) - E(y_2)]^2}}.$$

This formula indicates that the upper limit of the individual-level correlation is the correlation between variables, which is the correlation between the differences in the answers of the partners in any given dyad.

We can approximate this positive correlation by supposing that the expected values of the answers given by the two informants of any dyad or pair to the two questions/variables are equal:

$$r_i(X, X', Y, Y') = \frac{\sqrt{var(x_1 - x_2)} \cdot \sqrt{var(y_1 - y_2)} \cdot r(x_1 - x_2, y_1 - y_2) + [E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{\sqrt{var(x_1 - x_2) + [E(x_1) - E(x_2)]^2} \cdot \sqrt{var(y_1 - y_2) + [E(y_1) - E(y_2)]^2}}$$
$$\leq [r(x_1 - x_2, y_1 - y_2)].$$

This proves that this correlation actually measures the difference in the individual effect between dyads.

2.4 Dyad-level correlation

Lastly, we discuss the dyad-level correlation, which is described by the following formula:

$$r_d(X, X', Y, Y') = \frac{r(X, Y')}{\sqrt{r(X, X')} \cdot \sqrt{r(Y, Y')}}.$$

Let us remark that this is not a strict correlation in traditional statistical terms, since the variables under the square root might have negative values. This happens when the informants of a dyad give opposite answers to a question. Now, we set aside this problem and suppose that the expression under the square root is non-negative. The above formula can be transformed using the definition of correlation as follows using (1)–(6):

$$r_d(X, X', Y, Y') = \frac{\frac{1}{2} \cdot [\sqrt{var(x_1)} \cdot \sqrt{var(y_2)} \cdot r(x_1, y_2) + \sqrt{var(x_2)} \cdot \sqrt{var(y_1)} \cdot r(x_2, y_1)] - \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}}{\sqrt{cov(x_1, x_2) - \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \cdot \sqrt{cov(y_1, y_2) - \left(\frac{E(y_1) - E(y_2)}{2}\right)^2}}.$$

We can see that if:

$$cov(x_1, x_2) - \left(\frac{E(x_1) - E(x_2)}{2}\right)^2 < 0 \text{ and/or}$$
$$cov(y_1, y_2) - \left(\frac{E(y_1) - E(y_2)}{2}\right)^2 < 0,$$

then this type of correlation cannot be produced. This reflects that dyad-level correlation is similar to cross-intraclass correlation, as discussed in the context of homogeneity analysis.

When analyzing the covariance in the numerator of our expression, we can see that the correct correlation here is the cross-intraclass correlation $r(X, Y')$. This result can also be obtained by supposing the members of the dyads give similar answers to the questions. In such cases, the covariance becomes close to the variance because the expected values and standard deviations are close to each other.

2.5 Testing our suggested approximation formulas

In the previous sections, we critically analyzed five correlations, which were developed by dyadic data analysis. In the next table, we summarize these correlations, giving both the formulas developed by DDA and our suggested approximations.

These correlations were analyzed locally, and approximations were developed. We summarize our results in Table 7.

Table 6 DDA correlations with double entry and using the initial database

Type of correlation	(X, X'), (Y, Y') (double entry)	$(x_1, x_2), (y_1, y_2)$ (initial data)
Cross- intraclass correlation	$r(X, Y') = \frac{cov(X, Y')}{\sqrt{var(X)} \cdot \sqrt{var(Y)}}$,	$r(X, Y') = \frac{\frac{cov(x_1 - y_2) + cov(x_2 - y_1)}{2} - \frac{[E(x_1) - E(y_2)] \cdot [E(y_1) - E(y_2)]}{4}}{\sqrt{\frac{var(x_1) + var(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \cdot \sqrt{\frac{var(y_1) + var(y_2)}{2} + \left(\frac{E(y_1) - E(y_2)}{2}\right)^2}}$.
Overall within-partner correlation	$r(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)} \cdot \sqrt{var(Y)}}$,	$r(X, Y) = \frac{\frac{cov(x_1 - y_1) + cov(x_2 - y_2)}{2} + \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}}{\sqrt{\frac{var(x_1) + var(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \cdot \sqrt{\frac{var(y_1) + var(y_2)}{2} + \left(\frac{E(y_1) - E(y_2)}{2}\right)^2}}$.
Mean-level correlation	$r_m(X, X', Y, Y') = \frac{r(X, Y) + r(X, Y')}{\sqrt{1 + r(X, X')} \cdot \sqrt{1 + r(Y, Y')}}$,	$r_m(X, X', Y, Y') = r(x_1 + x_2, y_1 + y_2)$.
Individual-level correlation	$r_i(X, X', Y, Y') = \frac{r(X, Y) - r(X, Y')}{\sqrt{1 - r(X, X')} \cdot \sqrt{1 - r(Y, Y')}}$.	$r_i(X, X', Y, Y') = \frac{cov(x_1 - x_2, y_1 - y_2) + [E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{\sqrt{var(x_1 - x_2) + [E(x_1) - E(x_2)]^2} \cdot \sqrt{var(y_1 - y_2) + [E(y_1) - E(y_2)]^2}}$.
Dyad-level correlation	$r_d(X, X', Y, Y') = \frac{r(X, Y')}{\sqrt{r(X, X')} \cdot \sqrt{r(Y, Y')}}$,	$r_d(X, X', Y, Y') = \frac{\frac{cov(x_1, y_2) + cov(x_2, y_1)}{2} - \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}}{\sqrt{cov(x_1, x_2) - \left(\frac{E(x_1) - E(x_2)}{2}\right)^2} \cdot \sqrt{cov(y_1, y_2) - \left(\frac{E(y_1) - E(y_2)}{2}\right)^2}}$.

Source: Own construction

Table 7 Suggested approximations of the correlations specified by DDA using the initial, raw data

Types of DDA correlations	Suggested approximations
Overall within-partner correlation	$r(X, Y) \sim \frac{1}{2} \cdot [r(x_1, y_1) + r(x_2, y_2)]$
Cross-intraclass correlation	$r(X, Y') \sim \frac{1}{2} \cdot [r(x_1, y_2) + r(x_2, y_1)]$
Mean-level correlation	$r_m(X, X', Y, Y') = r(x_1 + x_2, y_1 + y_2)$
Individual-level correlation	$r_i(X, X', Y, Y') \sim r(x_1 - x_2, y_1 - y_2)$

Source: Own construction

We have theoretically elaborated the different correlation types and developed the formulas presented above. These formulas enable to avoid the application of double entry and to approximate correlations using the initial/raw data. Using our database and the same variables as before, we have calculated these correlations applying both the suggested traditional DDA formulas based on double entry and our developed expressions based on the initial data. Our objective is to test whether our suggested approximations lead to good results. If this is the case, the technique of double entry does not necessarily lead to additional information for statistical analysis. In Table 8, we have summarized the results of our empirical tests.

Table 8 Summary of the results of testing the correlation coefficients using the formulas of dyadic data analysis (based on double entry) and the suggested approximations (with initial database)

Types of DDA correlations	Values calculated using the database developed through double entry	Values calculated using the initial dataset
Overall within-partner correlation	0.291	0.293
Cross-intraclass correlation	0.588	0.589
Mean-level correlation	0.617	0.617
Individual-level correlation	0.522	0.522
Dyad-level correlation	0.689	0.293

Source: Own construction

Our suggested approximations resulted in good agreement with the original correlation indices of DDA, except for the dyad-level correlation. This result also supports our statement that the database development technique of double entry does not always yield significant benefit for statistical analysis.

3 DYADIC REGFRESSION MODELS

The core question of regression models is the effect that the independent variable has on the dependent variable. It is assumed that it is easier to specify independent variables in classical statistics compared to dyadic data analysis because DDA takes into account not only individual-level but also dyad-level effects. Therefore, regression analysis of dyadic data necessitates incorporating several factors, even if we have only one independent and one dependent variable. These factors are as follows (Gonzalez, 2010):

- Actor effect,
- Partner effect,
- Mutual effect.

The model of the intraclass correlation coefficient (ICC) incorporates only the actor and partner effects, while the actor-partner interdependence model (APIM) takes into consideration the mutual effect as well.

3.1 Theoretical discussion

In this section, we discuss the ICC model. First, we introduce the model. The objective is to analyze critically whether this linear model is capable of describing complex relationships between its dyadic variables. We know that the ICC model aims at describing only the actor and partner effects.

The model is formulated mathematically as follows (Gonzalez and Griffin, 2000):

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X' + \varepsilon ,$$

where X and X' are the independent variables that we obtained using double entry, Y is the dependent variable, ε is the error vector, and β_0, β_1 and β_2 are the regression coefficients.

This model can also be expressed in terms of the initial database as follows:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \beta_0 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \beta_1 \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \beta_2 \cdot \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix},$$

where vector 1 is a vector, in which all elements are equal to 1, and ε_1 and ε_2 are the error vectors. We unfold this estimation to examine its elements:

$$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon_1 \text{ and}$$

$$y_2 = \beta_0 \cdot 1 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1 + \varepsilon_2 .$$

We see that regression parameters in the second equation are the same as those in the first. This means that the estimate based on a database developed using the double-entry technique loosely approximates the value of any variable given by the second member of the dyad as y_2.

Based on the above argument, the following formulas lead to a better estimate:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \varepsilon_{11} \text{ and}$$

$$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \varepsilon_{21} .$$

Here, we must estimate six coefficients instead of three. The main complication is that the previous two estimation equations are transformed into two independent equations that are not linked by any joint coefficients; ε_{11} and ε_{21} are the error vectors. Although we discuss only exchangeable cases in the paper, the proposed estimations can also be useful for distinguishable ones.[4]

One can also see that the estimate suggested above leads to a smaller error and parameters can capture linear relationships more precisely (given, of course, that both models use the same estimation method)[5].

We assume that parameters (β_{01}, β_{11}, β_{21}) and (β_{02}, β_{12}, β_{22}) optimize our estimation functions that are defined as the least square functions, i.e. f_1 (β_{01}, β_{11}, β_{12}) and f_2 (β_{02}, β_{12}, β_{22}). Rao et al. (2008) and Grosz (2011) describe the solution procedure in their works. In this case, the estimation function of the first model –which is obtained using the same methodology, namely, the least-squares procedure– leads to

$$f_1(\beta_0, \beta_1, \beta_2) + f_2(\beta_0, \beta_2, \beta_1).$$

Because f_1 (β_{01}, β_{11}, β_{12}) and f_2 (β_{02}, β_{12}, β_{22}) utilize optimal coefficients, the following hold:

$$f_1(\beta_{01}, \beta_{11}, \beta_{12}) \le f_1(\beta_0, \beta_1, \beta_2) \text{ and}$$

$$f_2(\beta_{02}, \beta_{12}, \beta_{22}) \le f_2(\beta_0, \beta_2, \beta_1),$$

which means that:

$$f_1(\beta_{01}, \beta_{11}, \beta_{12}) + f_2(\beta_{02}, \beta_{12}, \beta_{22}) \le f_1(\beta_0, \beta_1, \beta_2) + f_2(\beta_0, \beta_2, \beta_1) = f(\beta_0, \beta_1, \beta_2).$$

We proved that the modified linear model using the initial dataset offers a better estimate than the original estimate suggested by DDA. We continue our discussion with the APIM model.

The APIM model differs from the ICC model with respect to the mutual effect. This model not only maps the interrelations between the partners of the dyads (the actor and partner effects) but also incorporates into the model the interrelations among different dyads. The mathematical formula is:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X' + \beta_3 \cdot \langle X \cdot X' \rangle + \varepsilon,$$

where β_0, β_1 and β_2 are defined as in the case of the ICC model, and ε again denotes the error. The only difference between the two formulas is that the mutual effect is incorporated into the model using the expression $\beta_3 \cdot \langle X \cdot X' \rangle$.

In this case, vector $\langle X \cdot X' \rangle$ is a new variable reflecting the joint, mutual effect of the partners in the same dyad on one partner's (called the actor) Y variable (or answer).

Again, we can express the model using the initial dataset in the following way:

$$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_1 \text{ and}$$

[4] Let us note that the basic objective of this paper is to critically analyze the statistical consequences of double entry, so we apply the classic regression models (Gonzalez and Griffin, 2000). Both ICC and APIM have been further developed. Discussion of these extended models is not in the focus of our paper (N.N., 2019.)

[5] Let us use the least-squares procedure for the estimation. In this case, the two equations we obtained are independent. The estimation functions obtained by the least-squares procedure are quadratic functions in the case of the first equation f_1, while for the second equation f_2 the parameters minimize the estimation functions, so we obtain the following inequalities: $f_1() \le f_1()$ and $f_2() \le f_2()$. Since the parameters of the least-squares procedure maximize R^2, the two equations will result in a slightly better estimate. We can make a similar argument in the case of maximum likelihood estimation.

$$y_2 = \beta_0 \cdot 1 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_2 .$$

Expression $\langle x_1 \cdot x_2 \rangle$ denotes the vector that is created by multiplying the elements of vectors x_1 and x_2.

In this case, the following new functions are suggested:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_{11} \text{ and}$$

$$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_{21} .$$

The considerations discussed in relation to the ICC model are relevant here as well. Consequently, the estimation functions suggested above are superior.

3.2 Testing the suggested estimation functions for the ICC model

We have tested the suggested estimation functions for the ICC model and carried out calculations using the DDA functions. *Y* is the dependent variable, and *X* and *X'* are the independent variables.

Using the ICC model, the value of *R* was 0.588 (Table 9). The model and the coefficient of variable *X* were significant, but the coefficient of *X'* was not.

Table 9 Results of the ICC model

R	R^2	Adjusted R^2	Standard error
0.588	.346	.338	1.220

Independent variables: X, X'

ANOVA table

Model	Sum of squares	df	Mean of sum of squares	F	Sig.
Regression	137.819	2	68.910	46.276	.000
Residual	260.591	175	1.489		
Sum	398.410	177			

Dependent variable: Y
Independent variables: X, X'

Coefficients

Model	Non standardized coefficients		Standardized coefficient	t	Sig.
	B	Std. error	Beta		
Constant	.761	.096		7.938	.000
X	.495	.059	.586	8.357	.000
X'	.004	.059	.004	.063	.950

Dependent variable: *Y*
Source: Own construction

As a next step, we applied the model to the initial database, as suggested previously:

$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \varepsilon_{11}$ and

$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \varepsilon_{21}$.

Recall that here we must estimate six coefficients, instead of the three for the original ICC model, and we must use two independent, separate estimation functions. Here, ε_{11} and ε_{21} are the errors.

We have calculated the two regression models. The results of Model 1 are summarized in Table 10, and the results of Model 2 are summarized in Table 11.

3.2.1 Model using the initial or raw dataset

Table 10 Results of the regression model between y1 and x1, x2, respectively – Model 1

Variables

Model	Independent variables
1	x_1, x_2

Dependent variable: y_1

Model	R	R^2	Adjusted R^2	Standard error
1	.583	.339	.324	1.170

Independent variables: x_1, x_2

ANOVA table

Model		Sum of squares	df	Mean of sum of squares	F	Sig.
1	Regression	60.481	2	30.241	22.096	.000
	Residual	117.699	86	1.369		
	Sum	178.180	88			

Dependent variable: y_1
Independent variables: x_1, x_2

Coefficients

Model		Non standardized coefficients		Standardized coefficient	t	Sig.
		B	Std. error	Beta		
1	Constant	.738	.130		5.672	.000
	x_1	.438	.079	.560	5.569	.000
	x_2	.035	.082	.043	.429	.669

Dependent variable: y_1
Source: Own construction

3.2.2 Model using the initial or raw database

Table 11 Results of the regression model between y_2 and x_1, x_2 – Model 2

Variables

Model	Independent variables
2	x_1, x_2

Dependent variable: y_2

Model	R	R^2	Adjusted R^2	Standard error
2	.599	.358	.343	1.282

Independent variables: x_1, x_2

ANOVA table

Model		Sum of squares	df	Mean of sum of squares	F	Sig.
2	Regression	78.907	2	39.453	24.010	.000
	Residual	141.318	86	1.643		
	Sum	220.225	88			

Dependent variable: y_2
Independent variables: x_1, x_2

Coefficients

Model		Non standardized coefficients		Standardized coefficient	t	Sig.
		B	Std. error	Beta		
2	Constant	.793	.143		5.562	.000
	x_1	–.028	.086	–.033	–.328	.744
	x_2	.558	.090	.614	6.192	.000

Dependent variable: y_2
Source: Own construction

The calculations presented above support our theoretical argument; in the case of ICC, we have obtained very similar results with the two suggested models that leave out the technique of double entry and use the initial database for analysis.

SUMMARY – CONCLUSION

Dyadic phenomena have become highly important not only in sociology and psychology but in a networked economy for economics and management studies. The paper critically discussed a relatively new statistical methodology that was developed for analyzing such dyadic problems, called dyadic data analysis. We had two objectives with our work. On the one hand, we critiqued the database development of DDA related to exchangeable cases and suggested an algorithm for solving the problem transforming

such a data set into a distinguishable one. On the other hand, we concentrated on double entry and its statistical consequences for classical dyadic correlations and regression analysis.

We concluded that an exchangeable case can be traced back to a distinguishable one with a relatively simple algorithm. Because of the symmetry of the partners' roles in any dyad, the number of potential databases is the exponential function of the number of dyads in the initial dataset. Therefore, in exchangeable cases, one has to look for a consensus in the way data are treated. We suggested applying a transformation of the initial data that eliminates this symmetry, such as summing and/or calculating the absolute values of the data differences.

The second focal issue of the paper was the double-entry technique. We analyzed whether this technique adds value through developing a richer information base or leads to information losses. Our examination revealed that double entry does not supply additional information compared to the initial database. Rather, it might lead to information losses, consequently making the statistical analysis less reliable.

We discussed the different correlation constructs of DDA, clarified their statistical content, and succeeded in tracing them back to the classical Pearson correlation. These correlation constructs also do not require the use of double entry. We reduced these correlations to a formula that uses the initial database to approximate them. This formula was carried out not only for the correlation constructs of DDA but also for its regression models. Statistical discussion revealed that in the ICC and APIM models, the double-entry method might make the estimates less reliable. The suggested regression models that use the simple initial database can achieve better estimation.

After we developed the new correlations and regression equations using the initial database, we carried out empirical analysis as well. We tested the suggested approximations for all correlation constructs and the ICC regression model with an empirical database. This database was developed in a previous field study using a trust-related questionnaire with pairwise sampling. In respect of the correlations we had mainly supporting results. Except for the dyad-level correlation, our suggested formulas resulted in good approximations. Results of the suggested two ICC regression models led to a slightly higher R^2, however differences were quite small. Empirical results support our theoretical argument in this respect too. We have to emphasize though, that other empirical databases might lead to different results, so further empirical research is needed in this respect.

ACKNOWLEDGMENT

The project is supported by the Hungarian Scientific Research Fund (OTKA), Project No. K 115542.

References

BURK, J. W., STEGLICH, C. E. G., SNIJDERS, T. A. B. Beyond dyadic interdependence: actor-oriented models for co-evolving social networks and individual behaviors. *International Journal of Behavioral Development*, 2007, Vol. 31, No. 4. pp. 397–404.

DENG, L. AND KE-HAI, Y. Multiple-group analysis for structural equation modeling with dependent samples. *Structural Equation Modeling: A Multidisciplinary Journal*, 2015, 22(4), pp. 552–567.

GELEI, A. AND DOBOS, I. Mutual trustworthiness as a governance mechanism in business relationships – A dyadic data analysis. *Acta Oeconomica*, 2016, 66(4), pp. 661–684.

GELEI, A. AND SUGÁR, A. The challenge of researching dyadic phenomena – the comparison of dyadic data analysis and traditional statistical methods. *Hungarian Statistical Review*, 2017, Special Number 21, pp. 78–100.

GONZALEZ, R. AND GRIFFIN, D. The correlational analysis of dyad-level data in the distinguishable case. *Personal Relationships*, 1999, 6(4), pp. 449–469.

GONZALEZ, R. AND GRIFFIN, D. On the Statistics of Interdependence: Treating Dyadic Data with Respect. In: ICKES, W. AND DUCK, S. eds. *The Social Psychology of Personal Relationships*, John Wiley and Sons, Ltd., 2000, pp. 181–213.

GONZALEZ, R. *Dyadic Data Analysis* [online]. University of Michigan. 2010. [cit. 2.5.2011] <http://www.cfs.purdue.edu/CFF/documents/Families_and_Health/purdue.pdf>.

GRIFFIN, D. AND GONZALEZ, R. Correlational Analysis of Dyad-Level Data in the Exchangeable Case. *Psychological Bulletin*, 1995, 118(3), pp. 430–439.

GROSZ, J. Identification of Influential Points in a Linear Regression Model [online]. *Statistika: Statistics and Economy Journal*, 2011, No. 1, pp. 71–77.

KENNY, D. A. *Dyadic Analysis* [online]. 2015. [cit. 26.1.2019] <http://davidakenny.net/dyad.htm>.

KENNY, D. A., KASHY, D. A., COOK, W. L. *Dyadic data Analysis.* New York, London: The Guilford Press, 2006.

LEDERMANN, T., MACHO, S., KENNY, D. A. Assessing mediation in dyadic data using the actor-partner interdependence model. *Structural Equation Modeling: A Multidisciplinary Journal*, 2011, 18(4), pp. 595–612.

LEDERMANN, T. AND KENNY, D. A. A toolbox with programs to restructure and describe dyadic data. *Journal of Social and Personal Relationships*, 2015, Vol. 32(8), pp. 997–1011. DOI: 10.1177/0265407514555273

MCARDLE, J. J. Dynamic but structural equation modeling of repeated measures data. In: NESSELROADE, J. R. AND CATTEL, R. B. eds. *Handbook of multivariate experimental psychology*, 2nd Ed. New York: Plenum, 1988, pp. 561–614.

N.N. *Dyadic Data Analysis* [online]. [cit. 11.3.2019] <https://www.mailman.columbia.edu/research/population-health-methods/dyadic-data-analysis>.

PEUGH, J. L., DILILLO, D., PANUZIO, J. Analyzing mixed-dyadic data using structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 2013, 20(2), pp. 314–337.

PLANALP, E. M., DU, H., BRAUNGART-RIEKER, J. M., WANG, L. Growth Curve Modeling to Studying Change: A Comparison of Approaches Using Longitudinal Dyadic Data With Distinguishable Dyads. *Structural Equation Modeling: A Multidisciplinary Journal*, 2017, 24(1), pp. 129–147.

RAO, R. C., TOUTENBURG, H., HEUMANN, C. *Linear Models and Generalizations: Least Squares and Alternatives.* Berlin: Springer, 2008.

WHITTAKER, T. A., BERETVAS, S. N., FALBO, T. Dyadic Curve-of-Factors Model: An Introduction and Illustration of a Model for Longitudinal Nonexchangeable Dyadic Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 2014, 21(2), pp. 303–317.