

Partial and Ecological Correlation: a Common Three-Term Covariance Decomposition

Renato Guseo

Received: date / Accepted: date

Abstract Let (X, Y, Z) be a trivariate statistical variable observed at individual level. We propose a three-term decomposition of covariance between variables X and Y conditionally on the effects induced by the existence of a variable Z . The three terms are called *residual covariance*, *covariance lack of fit* and *covariance fit*, respectively. *Partial covariance*, between X and Y after removing the linear effects of Z , $\sigma_Z(X, Y)$, is the sum of the first two terms while *ecological covariance*, between the two regression functions $\mu_X(Z)$ and $\mu_Y(Z)$, $\text{Cov}(\mu_X(Z), \mu_Y(Z))$, is the sum of the last two terms and, consequently, covariance lack of fit is the common additive term.

Simple examples are given in two contexts: in ecological fallacy problems arising in linear modelling with aggregate level analysis contrasted with partial correlation step-wise procedures performed at individual level and in the special case of a two-level nested model. Previous basic decomposition is extended to a multivariate-multiple framework. Distinction between descriptive and stochastic approaches is not essential.

Keywords covariance decomposition · partial covariance · ecological covariance · ecological fallacy · two-level nested model · multivariate covariance decomposition

This research was partially supported by Fondazione Cassa di Risparmio di Padova e Rovigo, Progetto di Eccellenza 2007: "Innovation Diffusion Processes: Differential Models, Agent-Based Frameworks and Forecasting Methods".

R. Guseo
University of Padua, Department of Statistical Sciences, via C. Battisti 241, 35100 Padua, Italy;
tel. ++39-049-8274146;
Fax: ++39-049-8274170;
E-mail: renato.guseo@unipd.it

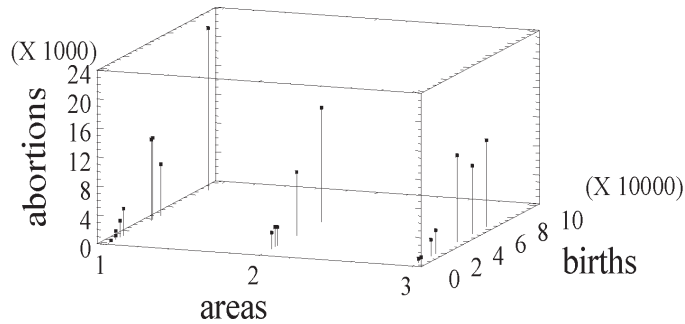


Fig. 1 Abortions and births in Italy at regional level, year 2005. Areas: “Nord” = 1, “Centro” = 2, “Sud” = 3. Data source: Italian Health Ministry.

1 Introduction

The Italian law n. 194 of May 2, 1978, introduced systematic regulations concerning maternity tutorship and legal abortion. The 2005 Health Ministry data at regional level and the corresponding mean values related to large geographic areas (“Nord”, “Centro” and “Sud”) highlight a contradictory behaviour. See, in particular, Figure 1. The linear correlation between legal number of abortions and births at regional level is $\rho_r = 0.95$. If we consider local mean values within large geographic areas we obtain correspondingly a negative relationship or, at least, a practical uncorrelation between mentioned factors, $\rho_a = -0.24$. This is a common contradictory situation emerging when comparing *aggregate* and *individual* level relationships that generates a well-known *ecological fallacy*. See, for instance, Robinson (1950), King et al. (2004), Faraway (2005).

The aim of the paper is to explain the analytical basis of such a paradox in linear model building by the means of a three-term decomposition of covariance.

Analysis of variance is a well-known statistical tool for the recognition of separate variability sources under the existence of an appropriate relationship between a dependent variable and corresponding covariates.

Let us consider, for simplicity, a bivariate statistical variable (X, Y) . The distinction between descriptive and probabilistic inferential approaches is not essential. In the discrete case the distinguishable outcomes from variable X are denoted by symbol x_i or, for simplicity, by subscript index i , $i = 1, 2, \dots, K_1$. Similarly, the separate outcomes from variable Y are denoted by symbol y_j or, by subscript index j , $j = 1, 2, \dots, K_2$.

In the discrete case, the joint relative frequency distribution or the corresponding probability joint distribution is represented by

$$p_{ij}, \quad i = 1, 2, \dots, K_1; \quad j = 1, 2, \dots, K_2.$$

Under a continuity hypothesis, especially within a probabilistic inferential framework, we denote by $p(x, y)$, $x \in X$, $y \in Y$ the joint probability density. In order to simplify exposition we use in the sequel the former notation. Limited changes are required in the continuous case.

Marginal distribution of Y can be represented by summing up, over all values of X , relative joint frequencies or joint probabilities under a fixed level y_j , i.e.,

$$p_{\cdot j} = \sum_{i=1}^{K_1} p_{ij}, \quad j = 1, 2, \dots, K_2.$$

One of the most important characterizing aspects of the marginal distribution of Y is based on conditional distributions of Y given $x \in X$, i.e., $Y|x$, $x \in X$. The *conditional* probability of y_j given level x_i , $p_{j|i}$, is defined as the ratio of the joint probability p_{ij} and the *conditioning* probability $p_{i\cdot}$,

$$p_{j|i} = \frac{p_{ij}}{p_{i\cdot}}, \quad p_{i\cdot} > 0.$$

A noticeable consequence is the characterization of the marginal distribution of Y as a *mixture* of conditional distributions or, in other words, an appropriate weighted mean of conditional distributions, i.e.,

$$p_{\cdot j} = \sum_{i=1}^{K_1} p_{ij} = \sum_{i=1}^{K_1} p_{j|i} p_{i\cdot}.$$

A similar property connects the marginal mean of Y , μ_Y , to the *regression function* of Y given x , i.e., $\mu_Y(x)$, $x \in X$, where $\mu_Y(x_i) = \sum_j y_j p_{j|i}$. It is easy to prove the following identity,

$$E(Y) = \mu_Y = E_X(\mu_Y(X)). \quad (1)$$

A different relationship links the marginal variance of Y , σ_Y^2 , to the *conditional variances* $\sigma_Y^2(x)$, $x \in X$, where $\sigma_Y^2(x_i) = \sum_j (y_j - \mu_Y(x_i))^2 p_{j|i}$. The well-known two-term variance decomposition highlights the joint role of conditional variances, $\sigma_Y^2(x)$, and regression function, $\mu_Y(x)$, i.e.,

$$\sigma_Y^2 = E_X(\sigma_Y^2(X)) + \bar{\sigma}_Y^2 = \sigma_Y^{*2} + \bar{\sigma}_Y^2, \quad (2)$$

where $\bar{\sigma}_Y^2 = E_X[(\mu_Y(X) - \mu_Y)^2]$ is the *explained variance* (by regression function) and σ_Y^{*2} denotes the *residual variance*. Previous decomposition is a special case of more general covariance decompositions (see, for instance, Guseo (2006), pp. 274–282).

Let us consider a trivariate statistical variable (X, Y, Z) observed at individual level. Covariance between X and Y , σ_{XY} , is a basic corner–stone in the analysis of linear models.

How can we analyze the information content of a covariance?

We know that the best predictor of Y given x , $Y|x$, determined according to a squared loss function, is the conditional mean, $\mu_Y(x)$, an *aggregate* notion that describes a general local property of a subpopulation characterized by $X = x$.

There is a well–known connection between $\mu_Y(x)$ and covariance σ_{XY} , i.e.,

$$\sigma_{XY} = E(XY) - E(X)E(Y) = E_X(X\mu_Y(X)) - E(X)E_X(\mu_Y(X)).$$

Therefore, covariance is a function of the regression function $\mu_Y(X)$ (or $\mu_X(Y)$). Nevertheless, it may be influenced by a concomitant variable Z . We analyze this further connection within the simplest trivariate case, (X, Y, Z) , and extend it in the final part of the paper.

Covariance may be decomposed into three additive terms in order to detect different sources in dependence analysis: *residual covariance*, σ_{XY}^* , *covariance lack of fit*, ${}_L\sigma_{XY}$, and first order *covariance fit*, ${}_C\sigma_{XY}$.

The aim of the paper is to prove that *partial covariance* between X and Y – after removing the linear effects of Z , $\sigma_Z(X, Y)$, which is the basic tool of step–wise model selection technique in linear models – is based on the sum of the first two components. Conversely, the direct comparison of the regression functions, $\mu_X(Z)$ and $\mu_Y(Z)$, based on the same conditioning factor Z , determines an *ecological covariance*, $\text{Cov}(\mu_X(Z), \mu_Y(Z))$, with well–known *ecological fallacy* problems depending on the aggregate nature of regression functions $\mu_X(z)$ and $\mu_Y(z)$, $z \in Z$. It will be proved that ecological covariance depends upon the sum of the last two terms in previous trivariate decomposition. The covariance lack of fit term is the common element of both statistical tools.

In this paper we are not interested in the so called “ecological inference problem” which is based on non–unique conditional distributions reconstruction in simple contingency tables with known marginals. A recent state of art monograph edited by King et al. (2004) illustrates the extensive work and advances in this special research area.

The paper is organized as follows. Section 2 presents a two–term decomposition of covariance, *residual covariance* and *ecological covariance*. Sections 3 and 4 present a three–term covariance decomposition based on ecological covariance splitting. This decomposition, see Sections 5 – 7, gives a useful insight into the relationship between *partial correlation* and *ecological correlation* coefficients. Section 8 examines the role of proposed three–term decomposition within a two–level hierarchical linear model. Section 9 presents a multivariate–multiple version of the three–term decomposition. Section 10 is left for final conclusions and remarks.

2 Two-term decomposition of covariance: residual and ecological components

The variance of variable Y , σ_Y^2 , may be influenced by the existence of a concomitant variable X . Note that individual observations define a bivariate variable (X, Y) . If we exclude stochastic independence or the weaker mean independence, this influence may be represented by a special shape of regression function. In particular, if residual variance is equal to zero, the variability of Y is due to the variations of X through regression function $\mu_Y(X)$.

We are often interested in evaluating covariance between the components of a bivariate variable (X, Y) under the effects of a concomitant variable Z . We assume here that individual observations define a trivariate variable (X, Y, Z) .

We suppose that a large influence of variable Z on covariance σ_{XY} may be explained through its regression functions, $\mu_X(Z)$ and $\mu_Y(Z)$.

Proposition 1. Let (X, Y) be a marginal bivariate statistical variable and Z a concomitant variable within (X, Y, Z) . Covariance σ_{XY} is the sum of two components: *residual covariance* σ_{XY}^* and *ecological covariance* $\sigma_{\bar{X}\bar{Y}}$,

$$\sigma_{XY} = \sigma_{XY}^* + \sigma_{\bar{X}\bar{Y}}, \quad (3)$$

where $\sigma_{XY}(Z) = E_{XY|Z}(X - \mu_X(Z))(Y - \mu_Y(Z))$, $\sigma_{XY}^* = E_Z\{\sigma_{XY}(Z)\}$ and $\sigma_{\bar{X}\bar{Y}} = \text{Cov}(\mu_X(Z), \mu_Y(Z))$.

Proof.

$$\begin{aligned} \sigma_{XY} &= E_Z E_{YX|Z}(X - \mu_X)(Y - \mu_Y) \\ &= E_Z E_{YX|Z}(X - \mu_X(Z) + \mu_X(Z) - \mu_X)(Y - \mu_Y(Z) + \mu_Y(Z) - \mu_Y) \\ &= E_Z\{\sigma_{XY}(Z)\} + E_Z E_{YX|Z}(X - \mu_X(Z))(\mu_Y(Z) - \mu_Y) + \\ &\quad + E_Z E_{YX|Z}(\mu_X(Z) - \mu_X)(Y - \mu_Y(Z)) + \\ &\quad + E_Z E_{YX|Z}(\mu_X(Z) - \mu_X)(\mu_Y(Z) - \mu_Y) \\ &= E_Z\{\sigma_{XY}(Z)\} + E_Z(\mu_X(Z) - \mu_X)(\mu_Y(Z) - \mu_Y) \\ &= E_Z\{\sigma_{XY}(Z)\} + \text{Cov}(\mu_X(Z), \mu_Y(Z)) = \sigma_{XY}^* + \sigma_{\bar{X}\bar{Y}}. \end{aligned}$$

□

It is well-known that covariance of a variable W with itself, σ_{WW} , denotes the variance of W , σ_W^2 . Therefore, decomposition in Equation (3) allows the corresponding usual variance decompositions with reference to the corresponding regression functions dependent on Z . We observe that *explained variance* is conceptually equivalent to *ecological variance*. As a matter of fact, for $X = Y$, we obtain

$$\begin{aligned} \sigma_{YY} &= \sigma_Y^2 = E_Z\{\sigma_{YY}(Z)\} + \text{Cov}(\mu_Y(Z), \mu_Y(Z)) \\ &= E_Z\{\sigma_Y^2(Z)\} + \text{Var}(\mu_Y(Z)) = \sigma_Y^{*2} + \sigma_{\bar{Y}}^2 = \sigma_Y^{*2} + \bar{\sigma}_{\bar{Y}}^2 \end{aligned} \quad (4)$$

and, for $Y = X$, we have

$$\sigma_{XX} = E_Z\{\sigma_X^2(Z)\} + \text{Var}(\mu_X(Z)) = \sigma_X^{*2} + \sigma_{\bar{X}}^2 = \sigma_X^{*2} + \bar{\sigma}_{\bar{X}}^2. \quad (5)$$

3 Two-term decomposition of ecological covariance

Ecological covariance represents the direct covariance between regression functions $\mu_X(Z)$ and $\mu_Y(Z)$. We study this relationship by introducing a deviation from the corresponding linear models, $X = a + bZ + \varepsilon_1$ and $Y = c + dZ + \varepsilon_2$. In the sequel, least squares parameters' estimates are denoted by a "hat".

Proposition 2. Let (X, Y) be a marginal bivariate statistical variable and Z a concomitant variable within (X, Y, Z) . Ecological covariance $\sigma_{\bar{X}\bar{Y}}$ is the sum of two components: the *covariance lack of fit*, ${}_L\sigma_{XY}$, and the first order *covariance fit*, ${}_C\sigma_{XY}$,

$$\sigma_{\bar{X}\bar{Y}} = {}_L\sigma_{XY} + {}_C\sigma_{XY}, \quad (6)$$

where ${}_L\sigma_{XY} = E_Z(\mu_X(Z) - \hat{a} - \hat{b}Z)(\mu_Y(Z) - \hat{c} - \hat{d}Z)$ and ${}_C\sigma_{XY} = E_Z(\hat{a} + \hat{b}Z - \mu_X)(\hat{c} + \hat{d}Z - \mu_Y)$.

Proof.

$$\begin{aligned} \sigma_{\bar{X}\bar{Y}} &= \text{Cov}(\mu_X(Z), \mu_Y(Z)) = E_Z(\mu_X(Z) - \mu_X)(\mu_Y(Z) - \mu_Y) \\ &= E_Z(\mu_X(Z) \pm (\hat{a} + \hat{b}Z) - \mu_X)(\mu_Y(Z) \pm (\hat{c} + \hat{d}Z) - \mu_Y) \\ &= E_Z(\mu_X(Z) - \hat{a} - \hat{b}Z)(\mu_Y(Z) - \hat{c} - \hat{d}Z) + \\ &\quad + E_Z(\hat{a} + \hat{b}Z - \mu_X)(\hat{c} + \hat{d}Z - \mu_Y) \\ &= E_Z(\mu_X(Z) - \hat{a} - \hat{b}Z)(\mu_Y(Z) - \hat{c} - \hat{d}Z) + \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma_Z^2} \\ &= {}_L\sigma_{XY} + {}_C\sigma_{XY}, \end{aligned}$$

where the reduced form of the last term, the *covariance fit*, is determined as follows,

$$\begin{aligned} {}_C\sigma_{XY} &= E_Z(\hat{a} + \hat{b}Z - \mu_X)(\hat{c} + \hat{d}Z - \mu_Y) \\ &= E_Z(\mu_X - \hat{b}\mu_Z + \hat{b}Z - \mu_X)(\mu_Y - \hat{d}\mu_Z + \hat{d}Z - \mu_Y) \\ &= \hat{b}\hat{d}E(Z - \mu_Z)^2 = \hat{b}\hat{d}\sigma_Z^2 = \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma_Z^2}. \end{aligned} \quad (7)$$

□

If at least one of the two regression functions, $\mu_X(Z)$ or $\mu_Y(Z)$, is essentially linear, then covariance lack of fit is equal to zero, ${}_L\sigma_{XY} = 0$, and ecological covariance is $\sigma_{\bar{X}\bar{Y}} = \sigma_{XZ}\sigma_{YZ}/\sigma_Z^2$.

Because of Equations (6) and (7) we can obtain a parallel decomposition of *explained variances* $\bar{\sigma}_X^2 = \sigma_{\bar{X}\bar{X}} = \sigma_X^2$ and $\bar{\sigma}_Y^2 = \sigma_{\bar{Y}\bar{Y}} = \sigma_Y^2$, i.e.,

$$\bar{\sigma}_X^2 = {}_L\sigma_X^2 + {}_C\sigma_X^2, \quad (8)$$

where ${}_L\sigma_X^2 = E_Z(\mu_X(Z) - \hat{a} - \hat{b}Z)^2$ and ${}_C\sigma_X^2 = \sigma_{XZ}^2/\sigma_Z^2$.

Similarly, we obtain

$$\bar{\sigma}_Y^2 = {}_L\sigma_Y^2 + {}_C\sigma_Y^2, \quad (9)$$

where ${}_L\sigma_Y^2 = E_Z(\mu_Y(Z) - \hat{c} - \hat{d}Z)^2$ and ${}_C\sigma_Y^2 = \sigma_{YZ}^2/\sigma_Z^2$.

4 Three-term decomposition of covariance and partial covariance

A simple combination of Propositions 1 and 2 yields a three-term decomposition of covariance.

Proposition 3. Let (X, Y) be a marginal bivariate statistical variable and Z a concomitant variable in (X, Y, Z) . Covariance between X and Y can be decomposed as follows:

$$\sigma_{XY} = \sigma_{XY}^* + {}_L\sigma_{XY} + {}_C\sigma_{XY}, \quad (10)$$

where $\sigma_{XY}^* = E_Z\{\sigma_{XY}(Z)\}$ is the *residual covariance*, ${}_L\sigma_{XY} = E_Z(\mu_X(Z) - \hat{a} - \hat{b}Z)(\mu_Y(Z) - \hat{c} - \hat{d}Z)$ is the *covariance lack of fit* and ${}_C\sigma_{XY} = E_Z(\hat{a} + \hat{b}Z - \mu_X)(\hat{c} + \hat{d}Z - \mu_Y) = (\sigma_{XZ}\sigma_{YZ})/\sigma_Z^2$ is the *first order covariance fit*.

By Proposition 1, the sum of the last two terms in Equation (10) represents *ecological covariance*. Moreover, the sum of the first two components in Equation (10) corresponds to a well-known statistical concept.

Proposition 4. Let (X, Y) be a marginal bivariate statistical variable and Z a concomitant variable in (X, Y, Z) . *Partial covariance* between X and Y after removing the linear effects of Z , $\sigma_Z(X, Y) = E[(X - \hat{a} - \hat{b}Z)(Y - \hat{c} - \hat{d}Z)]$, is the sum of the first two components in Equation (10), i.e., residual covariance and covariance lack of fit,

$$\sigma_Z(X, Y) = \sigma_{XY}^* + {}_L\sigma_{XY}. \quad (11)$$

Proof.

We introduce in partial covariance the deviations with respect to the corresponding regression functions, $\mu_X(Z)$ and $\mu_Y(Z)$, so that we obtain,

$$\begin{aligned} \sigma_Z(X, Y) &= E[(X \pm \mu_X(Z) - \hat{a} - \hat{b}Z)(Y \pm \mu_Y(Z) - \hat{c} - \hat{d}Z)] \\ &= E_Z\{\sigma_{XY}(Z)\} + E_Z[(\mu_X(Z) - \hat{a} - \hat{b}Z)(\mu_Y(Z) - \hat{c} - \hat{d}Z)] \\ &= \sigma_{XY}^* + {}_L\sigma_{XY}. \end{aligned}$$

□

5 Direct, ecological and partial correlation coefficients

On the basis of previous results we may define three different squared correlation coefficients according to the following notations,

$$\sigma_X^{*2} = E_Z(\sigma_X^2(Z)), \quad {}_C\sigma_X^2 = \frac{\sigma_{XZ}^2}{\sigma_Z^2}, \quad {}_L\sigma_X^2 = E_Z(\mu_X(Z) - \hat{a} - \hat{b}Z)^2$$

and

$$\sigma_Y^{*2} = E_Z(\sigma_Y^2(Z)), \quad {}_C\sigma_Y^2 = \frac{\sigma_{YZ}^2}{\sigma_Z^2}, \quad {}_L\sigma_Y^2 = E_Z(\mu_Y(Z) - \hat{c} - \hat{d}Z)^2.$$

Note that for $X = Y$ we obtain from Equation (11) the corresponding *partial variances*, i.e.,

$$\sigma_Z^2(X) = \sigma_X^{*2} + L\sigma_X^2, \quad \sigma_Z^2(Y) = \sigma_Y^{*2} + L\sigma_Y^2. \quad (12)$$

Squared direct correlation coefficient is

$$\rho_{XY}^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \frac{(\sigma_{XY}^* + L\sigma_{XY} + C\sigma_{XY})^2}{(\sigma_X^{*2} + L\sigma_X^2 + C\sigma_X^2)(\sigma_Y^{*2} + L\sigma_Y^2 + C\sigma_Y^2)}, \quad (13)$$

squared ecological correlation coefficient is

$$\rho_{\bar{X}\bar{Y}}^2 = \frac{\sigma_{\bar{X}\bar{Y}}^2}{\sigma_{\bar{X}}^2 \sigma_{\bar{Y}}^2} = \frac{(L\sigma_{XY} + C\sigma_{XY})^2}{(L\sigma_X^2 + C\sigma_X^2)(L\sigma_Y^2 + C\sigma_Y^2)}, \quad (14)$$

and, finally, squared partial correlation coefficient is

$${}_{XY}\rho_Z^2 = \frac{\sigma_Z^2(X, Y)}{\sigma_Z^2(X)\sigma_Z^2(Y)} = \frac{(\sigma_{XY}^* + L\sigma_{XY})^2}{(\sigma_X^{*2} + L\sigma_X^2)(\sigma_Y^{*2} + L\sigma_Y^2)}. \quad (15)$$

Following the three-term decomposition in Equation (10), *partial covariance* between X and Y , after removing the linear effects of Z , is

$$\sigma_Z(X, Y) = \sigma_{XY} - \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma_Z^2} = \sigma_{XY} - C\sigma_{XY}. \quad (16)$$

Notice that if at least one of the two regression functions, $\mu_X(Z)$ or $\mu_Y(Z)$, is linear, then $\sigma_{XY} = \sigma_{XY}^* + C\sigma_{XY}$ and therefore, partial covariance and ecological covariance have simpler forms since the *covariance lack of fit* is zero, $L\sigma_{XY} = 0$.

Moreover, if both regression functions are linear, we obtain

$${}_{XY}\rho_Z^2 = \frac{\sigma_{XY}^{*2}}{\sigma_X^{*2}\sigma_Y^{*2}}; \quad \rho_{\bar{X}\bar{Y}}^2 = 1; \quad \rho_{XY}^2 = \frac{(\sigma_{XY}^* + C\sigma_{XC}\sigma_Y)^2}{(\sigma_X^{*2} + C\sigma_X^2)(\sigma_Y^{*2} + C\sigma_Y^2)}.$$

If Z is uncorrelated with both X and Y then we have $C\sigma_{XY} = 0 = C\sigma_X^2 = C\sigma_Y^2$, and therefore, $\rho_{\bar{X}\bar{Y}}^2 = {}_{XY}\rho_Z^2$.

6 Partial correlation and nested linear models

To build an efficient model within a multiple regressive framework, *partial covariance* is more important than ecological covariance. In particular, *partial covariance* is the major tool in parsimonious model selection according to nonparametric backward step-wise procedures. Let us consider the most simple case concerning squared partial correlation between variables X and Y conditionally on a concomitant variable Z . In Section 5, see Equation (16), we determined partial covariance or, equivalently, the covariance of appropriate residual variables after removing the linear effect of

variable Z . Let us define the corresponding residual variances, following an alternative formulation, i.e.,

$$\begin{aligned}\sigma_Z^2(X) &= E[(X - \hat{a} - \hat{b}Z)^2] = E[(X - \mu_X + \hat{b}\mu_Z - \hat{b}Z)^2] \\ &= E[(X - \mu_X) - \hat{b}(Z - \mu_Z)]^2 = \sigma_X^2 + \frac{\sigma_{XZ}^2}{\sigma_Z^4}\sigma_Z^2 - 2\frac{\sigma_{XZ}^2}{\sigma_Z^2} \\ &= \sigma_X^2 - \frac{\sigma_{XZ}^2}{\sigma_Z^2} = \sigma_X^2(1 - \rho_{XZ}^2)\end{aligned}\quad (17)$$

and

$$\sigma_Z^2(Y) = E[(Y - \hat{c} - \hat{d}Z)^2] = \sigma_Y^2 - \frac{\sigma_{YZ}^2}{\sigma_Z^2} = \sigma_Y^2(1 - \rho_{YZ}^2). \quad (18)$$

The squared partial correlation coefficient between X and Y after removing the linear effects due to variable Z is a well-known result,

$${}_{XY}\rho_Z^2 = \frac{(\sigma_{XY}\sigma_Z^2 - \sigma_{XZ}\sigma_{YZ})^2/\sigma_Z^4}{(\sigma_Y^2\sigma_Z^2 - \sigma_{YZ}^2)(\sigma_X^2\sigma_Z^2 - \sigma_{XZ}^2)/\sigma_Z^4} = \frac{(\rho_{XY} - \rho_{XZ}\rho_{YZ})^2}{(1 - \rho_{YZ}^2)(1 - \rho_{XZ}^2)}. \quad (19)$$

If squared partial correlation coefficient is zero, ${}_{XY}\rho_Z^2 = 0$, we attain $\rho_{XY} = \rho_{XZ}\rho_{YZ}$. Alternatively, by Equation (10), we note that marginal covariance between X and Y is,

$$\sigma_{XY} = \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma_Z^2},$$

or, equivalently,

$$\rho_{XY} = \rho_{XZ}\rho_{YZ}. \quad (20)$$

Apparent informative contribution (spurious effect) of variable X on Y in σ_{XY} or in ρ_{XY} explicitly depends on the presence of non-zero linear ties between Z and X and, respectively, between Z and Y . However, the linear incremental informative contribution of X , in a linear regressive model containing Z as regressor and Y as a dependent variable, is perfectly useless if ${}_{XY}\rho_Z^2 = 0$.

It is well-known that squared partial correlation equivalently denotes the relative variation of residual variance between two linear *nested models*, i.e.,

$${}_{XY}\rho_Z^2 = \frac{{}_Y\sigma_Z^{*2} - {}_Y\sigma_{ZX}^{*2}}{{}_Y\sigma_Z^{*2}},$$

where ${}_Y\sigma_{ZX}^{*2}$ is the residual variance in the multiple regressive model $Y = \alpha_0 + \alpha_1 Z + \alpha_2 X + \varepsilon$, while ${}_Y\sigma_Z^{*2} = \sigma_Y^2(1 - \rho_{ZY}^2)$ is the residual variance of the simpler regression line, $Y = \alpha_0 + \alpha_1 Z + \varepsilon'$. If squared partial correlation coefficient is zero, ${}_{XY}\rho_Z^2 = 0$, then the corresponding residual variances are equal, ${}_Y\sigma_Z^{*2} = {}_Y\sigma_{ZX}^{*2}$; in other terms, the least squares estimation of the regressive coefficient relating to X is zero, $\hat{\alpha}_2 = 0$.

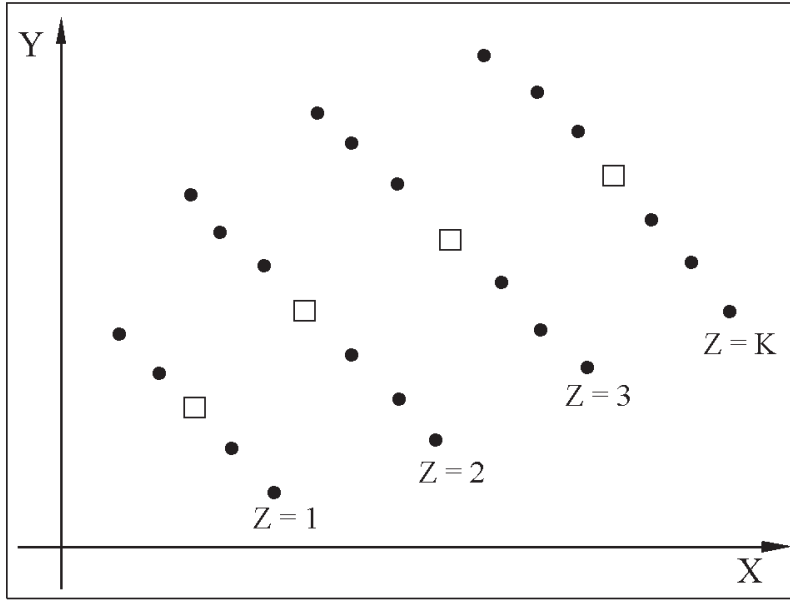


Fig. 2 Local negative relationships between Y and X for a fixed value $Z = 1, 2, \dots, K$ and ecological fallacy induced by positive relationship between $\mu_Y(Z)$ and $\mu_X(Z)$. Graphical symbols: • individual data, □ local mean values ($\mu_X(Z), \mu_Y(Z)$), $Z = 1, 2, \dots, K$.

7 Partial correlation and ecological correlation: a comparison

Let us consider a simple case where the relationship between two variables Y and X is monotonically decreasing conditionally on Z . See for instance the corresponding plot in Figure 2. Points represent individual observations while boxes denote local mean values of X and Y conditionally on Z , $(\mu_X(z), \mu_Y(z))$, $z = 1, 2, \dots, K$ which are monotonically increasing in (X, Y) space.

A positive value of ecological covariance, $\sigma_{\bar{X}\bar{Y}} = \text{Cov}(\mu_X(Z), \mu_Y(Z)) > 0$, and the corresponding positive value of ecological correlation coefficient, $\rho_{\bar{X}\bar{Y}} > 0$, may generate the well-known *ecological fallacy* typically due to an error in inferential reference.

Positive relationship between \bar{Y} and \bar{X} in the aggregate units space contrasts with the conditional relationship referred to the individual units. The latter relationship is essentially negative in sign. In this simple example if we assume that regression functions $\mu_X(Z)$ and $\mu_Y(Z)$ are essentially linear (affine transformations of Z), an appealing reduction arises, $L\sigma_{XY} = L\sigma_X^2 = L\sigma_Y^2 = 0$, so that ${}_{XY}\rho_Z = \sigma_{XY}^*/\sigma_X^*\sigma_Y^* < 0$ correctly denotes strength and sign (negative) of individual level relationship contrasting with aggregate level for which $\rho_{\bar{X}\bar{Y}} = \frac{c\sigma_{XY}}{|c\sigma_X c\sigma_Y|} > 0$ is positive.

Ecological fallacy represents a concrete problem if we use aggregate information and pretend to refer such a global relationship to individual units within special

subgroups. This problem was examined in the past in epidemiological and sociological areas. Robinson (1950), among others, is a well-known reference. New advances and critical discussions may be found in Freedman et al. (1991), Achen and Shilvely (1995), King (1997), Freedman et al. (1998) and Schuessler (1999).

More recently, in the monograph edited by King et al. (2004) there are special contributions to the “ecological inference problem” often focused on inferring internal cell counts in $r \times c$ contingency tables (usually 2×2 tables) from known marginal totals. Some interesting suggestions are presented in Chapters 10 – 12 by recognizing that the information omission regarding spatial aggregation of individuals into areal regions may cause specification bias in ecological inference.

In this paper we emphasize the role of three-term decomposition of covariance in the light of complex linear model building based on relevant covariates. Unfortunately, in some cases ecological fallacy, depending on aggregate data modelling, can not be avoided. See, for instance, the simple example stylized in Figure 2. If we study only individual data the paradox does not exist.

Moreover, the use of reduction procedure such as *step-wise backward* elimination of redundant components in a multiple regressive individual data framework allows the identification of a parsimonious model that avoids a risky reference to ecological covariance.

Let us consider again, as an explanatory description, the example related to Figure 2. Let (X, Z) be a statistical bivariate variable and Y a linear combination,

$$Y = \alpha + \beta X + \gamma Z, \quad \beta < 0, \gamma > 0.$$

Let us compute preliminarily covariances σ_{XY} and σ_{ZY} ,

$$\sigma_{XY} = \beta\sigma_X^2 + \gamma\sigma_{XZ}, \quad \sigma_{ZY} = \beta\sigma_{XZ} + \gamma\sigma_Z^2.$$

The determination of partial covariance is straightforward,

$$\begin{aligned} \sigma_Z(X, Y) &= \sigma_{XY} - \frac{\sigma_{XZ}\sigma_{ZY}}{\sigma_Z^2} = \beta\sigma_X^2 + \gamma\sigma_{XZ} - \frac{\sigma_{XZ}(\beta\sigma_{XZ} + \gamma\sigma_Z^2)}{\sigma_Z^2} \\ &= \beta\sigma_X^2 + \gamma\sigma_{XZ} - \beta\frac{\sigma_{XZ}^2}{\sigma_Z^2} - \gamma\sigma_{XZ} = \beta\left(\sigma_X^2 - \frac{\sigma_{XZ}^2}{\sigma_Z^2}\right) \\ &= \beta(1 - \rho_{XZ}^2)\sigma_X^2. \end{aligned} \quad (21)$$

Observe that $\mu_Y(Z) = E(Y|Z) = \alpha + \beta\mu_X(Z) + \gamma Z$. If $\mu_Y(Z)$ is an affine form depending on Z we obtain $\mu_Y(Z) = A + BZ$, and, therefore, ${}_L\sigma_{XY} = 0$.

Following this further hypothesis, ecological covariance has a simpler form and a direct computation of ${}_C\sigma_{XY}$ is sufficient,

$$\begin{aligned} \sigma_{\bar{X}\bar{Y}} &= {}_C\sigma_{XY} = \frac{\sigma_{XZ}\sigma_{ZY}}{\sigma_Z^2} = \frac{\sigma_{XZ}(\beta\sigma_{XZ} + \gamma\sigma_Z^2)}{\sigma_Z^2} \\ &= \beta\frac{\sigma_{XZ}^2}{\sigma_Z^2} + \gamma\sigma_{XZ} = \beta\rho_{XZ}^2\sigma_X^2 + \gamma\sigma_{XZ}. \end{aligned} \quad (22)$$

If we can assert that $\sigma_{XY} > 0$, a natural assumption into the framework represented in Figure 2, then, $\sigma_Z(X, Y) < 0$ because $\beta < 0$ and, moreover, $\sigma_{\bar{X}\bar{Y}} > 0$.

We underline, once again, the central role of partial correlation with reference to a weaker, sometimes misleading, ecological correlation.

If we apply, as a control check, least squares method to the exact model,

$$Y = \alpha + \beta X + \gamma Z, \quad \beta < 0, \gamma > 0,$$

we obtain,

$$\hat{\beta} = \frac{\sigma_{XY} \sigma_Z^2 - \sigma_{ZY} \sigma_{XZ}}{\sigma_X^2 \sigma_Z^2 - \sigma_{XZ}^2} = \beta,$$

$$\hat{\gamma} = \frac{\sigma_{ZY} \sigma_X^2 - \sigma_{XZ} \sigma_{XY}}{\sigma_X^2 \sigma_Z^2 - \sigma_{XZ}^2} = \gamma,$$

where ${}_{XY}\rho_Z^2 = 1$.

8 A two-level model

The proposed three-term decomposition of covariance is a useful tool in explaining special properties of a two-level linear model.

Let us consider the following model specification,

$$y_{ij} = \gamma_0 + \gamma_1 \tilde{x}_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, r; \quad j = 1, \dots, k \quad (23)$$

where: j denotes the reference group among k and i depicts a generic unit within the group characterized by a constant size, r . We introduce conditional X mean values with reference to j -th group, \bar{x}_j , i.e.,

$$y_{ij} = \gamma_0 + \gamma_1 (\tilde{x}_{ij} - \bar{x}_j) + \gamma_1 \bar{x}_j + \varepsilon_{ij},$$

and denoting by x_{ij} the deviation from corresponding local mean, $x_{ij} = \tilde{x}_{ij} - \bar{x}_j$, we obtain,

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \gamma_1 \bar{x}_j + \varepsilon_{ij}.$$

A generalization of the previous model is based on a free specification of the coefficient γ_2 related to the average value of X within group j -th, \bar{x}_j , i.e.,

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \gamma_2 \bar{x}_j + \varepsilon_{ij}. \quad (24)$$

In order to apply the least squares method we compute the appropriate variances and covariances, i.e.,

$$\sigma_X^2 = \frac{1}{rk} \sum_{ij} x_{ij}^2; \quad \sigma_{\bar{X}}^2 = \frac{1}{k} \sum_j (\bar{x}_j - \bar{x})^2,$$

$$\sigma_{\bar{X}Y} = \frac{1}{rk} \sum_{ij} (\bar{x}_j - \bar{x})(y_{ij} - \bar{y}) = \frac{1}{k} \sum_j (\bar{x}_j - \bar{x})(\bar{y}_j - \bar{y}) = \sigma_{\bar{X}\bar{Y}},$$

$$\sigma_{X\bar{X}} = \frac{1}{rk} \sum_{ij} x_{ij}(\bar{x}_j - \bar{x}) = 0. \quad (25)$$

By applying well-known formulas we obtain,

$$\begin{aligned}\hat{\gamma}_1 &= \frac{\sigma_{XY} \sigma_{\bar{X}}^2 - \sigma_{\bar{X}Y} \sigma_{X\bar{X}}}{\sigma_X^2 \sigma_{\bar{X}}^2 - \sigma_{X\bar{X}}^2} = \frac{\sigma_{XY}}{\sigma_X^2}, \\ \hat{\gamma}_2 &= \frac{\sigma_X^2 \sigma_{\bar{X}Y} - \sigma_{X\bar{X}} \sigma_{XY}}{\sigma_X^2 \sigma_{\bar{X}}^2 - \sigma_{X\bar{X}}^2} = \frac{\sigma_{\bar{X}Y}}{\sigma_{\bar{X}}^2} = \\ &= \frac{\sigma_{\bar{X}\bar{Y}}}{\sigma_{\bar{X}}^2} = \rho_{\bar{X}\bar{Y}} \frac{\sigma_{\bar{Y}}}{\sigma_{\bar{X}}}.\end{aligned}\quad (26)$$

Some attention is necessary in these cases in order to avoid wrong interpretations. Two-term decomposition of covariance σ_{XY} (see Equation (3)) gives rise to

$$\sigma_{XY} = \sigma_{XY}^* + \text{Cov}(\mu_X(\bar{X}), \mu_Y(\bar{X}))$$

where $\mu_X(\bar{X}) = 0$ uniformly by definition and then ecologic covariance is zero. Therefore, we obtain $\sigma_{XY} = \sigma_{XY}^*$ and $\sigma_{X\bar{X}} = \sigma_{X\bar{X}}^*$. The estimate of parameter γ_1 is then $\hat{\gamma}_1 = \sigma_{XY}^*/\sigma_X^2$.

The analysis of ecological covariance components is straightforward. We observe that *covariance fit* is also zero,

$$C\sigma_{XY} = \frac{\sigma_{X\bar{X}} \sigma_{Y\bar{X}}}{\sigma_{\bar{X}}^2} = \frac{0 \cdot \sigma_{\bar{Y}\bar{X}}}{\sigma_{\bar{X}}^2} = 0,$$

and therefore, ${}_L\sigma_{XY} = 0$.

Finally, we deduce that partial covariance is equal to residual covariance,

$$\sigma_{\bar{X}}(X, Y) = \sigma_{XY}^*, \quad (27)$$

and the squared partial correlation coefficient is

$$\begin{aligned}{}_{XY}\rho_{\bar{X}}^2 &= \frac{(\sigma_{XY} \sigma_{\bar{X}}^2 - 0 \cdot \sigma_{\bar{Y}\bar{X}})^2}{(\sigma_Y^2 \sigma_{\bar{X}}^2 - \sigma_{\bar{X}Y}^2)(\sigma_X^2 \sigma_{\bar{X}}^2 - 0)} \\ &= \frac{\sigma_{XY}^2 \sigma_{\bar{X}}^2}{(\sigma_Y^2 \sigma_{\bar{X}}^2 - \sigma_{\bar{X}Y}^2) \sigma_X^2} \\ &= \frac{\sigma_{\bar{X}Y}^2 \sigma_X^2}{\sigma_Y^2 \sigma_{\bar{X}}^2 (1 - \rho_{\bar{X}Y}^2) \sigma_X^2} \\ &= \frac{\rho_{\bar{X}Y}^2}{(1 - \rho_{\bar{X}Y}^2)}.\end{aligned}\quad (28)$$

Equivalence between ${}_{XY}\rho_{\bar{X}}^2$ and $\rho_{\bar{X}Y}^2$ is possible if and only if \bar{X} and Y are uncorrelated. Otherwise, squared partial correlation coefficient is greater than squared direct correlation,

$${}_{XY}\rho_{\bar{X}}^2 \geq \rho_{\bar{X}Y}^2.$$

9 Multivariate–Multiple Extension

Previous three–fold decomposition, summarized by Equations (10), may be extended to a multivariate variable with reference to a conditioning vector.

Let us define a statistical vector $\mathbf{X} \in \mathbb{R}^k$, $k > 1$ with mean vector $\mu_{\mathbf{X}} = E(\mathbf{X})$ and covariance matrix $\Sigma_{\mathbf{X}} = E[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})']$ and a concomitant statistical variable (or vector) Z . We consider a conditional distribution $\mathbf{X}|Z$ characterized by the conditional mean vector $\mu_{\mathbf{X}}(Z)$, i.e., the multivariate regression function of Z , and the conditional covariance matrix $\Sigma_{\mathbf{X}}(Z) = E_{\mathbf{X}|Z}[(\mathbf{X} - \mu_{\mathbf{X}}(Z))(\mathbf{X} - \mu_{\mathbf{X}}(Z))']$ so that we can define the *residual covariance matrix*

$$\Sigma_{\mathbf{X}}^* = E_Z(\Sigma_{\mathbf{X}}(Z)). \quad (29)$$

It is well–known that marginal mean vector $\mu_{\mathbf{X}}$ is related to the corresponding regression function $\mu_{\mathbf{X}}(Z)$, i.e.,

$$\mu_{\mathbf{X}} = E_Z(\mu_{\mathbf{X}}(Z)). \quad (30)$$

We may define a special covariance notion, the *ecological covariance matrix*,

$$\Sigma_{\bar{\mathbf{X}}} = E_Z[(\mu_{\mathbf{X}}(Z) - \mu_{\mathbf{X}})(\mu_{\mathbf{X}}(Z) - \mu_{\mathbf{X}})']. \quad (31)$$

Proposition 1. Let \mathbf{X} be a statistical vector and Z a concomitant variable. Covariance matrix $\Sigma_{\mathbf{X}}$ is the sum of two matrix components: *residual covariance matrix*, $\Sigma_{\mathbf{X}}^*$, and *ecological covariance matrix* $\Sigma_{\bar{\mathbf{X}}}$,

$$\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X}}^* + \Sigma_{\bar{\mathbf{X}}}. \quad (32)$$

Proof.

$$\begin{aligned} \Sigma_{\mathbf{X}} &= E_Z E_{\mathbf{X}|Z}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})'] \\ &= E_Z E_{\mathbf{X}|Z}[(\mathbf{X} - \mu_{\mathbf{X}}(Z) + \mu_{\mathbf{X}}(Z) - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}}(Z) + \mu_{\mathbf{X}}(Z) - \mu_{\mathbf{X}})'] \\ &= \Sigma_{\mathbf{X}}^* + \Sigma_{\bar{\mathbf{X}}}. \end{aligned}$$

Ecological covariance matrix describes the covariance of multivariate regression function $\mu_{\mathbf{X}}(Z)$. We may decompose such a relationship by introducing a deviation from the corresponding linear models determined under least squares method, e.g., $\mathbf{X} = A\mathbf{Z} + \varepsilon$, where $\mathbf{Z} = (1, Z)'$ and A is a suitable real matrix. Least squares solution is denoted by \hat{A} .

Proposition 2. Let \mathbf{X} be a statistical vector and Z a concomitant variable. The ecological covariance matrix $\Sigma_{\bar{\mathbf{X}}}$ is the sum of two components: the *covariance lack of fit* matrix, ${}_L\Sigma_{\bar{\mathbf{X}}}$, and the first order *covariance fit* matrix, ${}_C\Sigma_{\bar{\mathbf{X}}}$, i.e.,

$$\Sigma_{\bar{\mathbf{X}}} = {}_L\Sigma_{\bar{\mathbf{X}}} + {}_C\Sigma_{\bar{\mathbf{X}}}. \quad (33)$$

Proof.

$$\begin{aligned}
\Sigma_{\bar{\mathbf{X}}} &= E_Z[(\mu_{\mathbf{X}}(Z) - \mu_{\mathbf{X}})(\mu_{\mathbf{X}}(Z) - \mu_{\mathbf{X}})'] \\
&= E_Z[(\mu_{\mathbf{X}}(Z) \pm \hat{\mathbf{A}}\mathbf{Z} - \mu_{\mathbf{X}})(\mu_{\mathbf{X}}(Z) \pm \hat{\mathbf{A}}\mathbf{Z} - \mu_{\mathbf{X}})'] \\
&= E_Z[(\mu_{\mathbf{X}}(Z) - \hat{\mathbf{A}}\mathbf{Z})(\mu_{\mathbf{X}}(Z) - \hat{\mathbf{A}}\mathbf{Z})'] + E_Z[(\hat{\mathbf{A}}\mathbf{Z} - \mu_{\mathbf{X}})(\hat{\mathbf{A}}\mathbf{Z} - \mu_{\mathbf{X}})'] \\
&= {}_L\Sigma_{\mathbf{X}} + {}_C\Sigma_{\mathbf{X}}.
\end{aligned}$$

The cross-product mean values are zero because $\mu_{\mathbf{X}}(Z)$ and $\hat{\mathbf{A}}$ satisfy least squares principle and, therefore, also the corresponding normal equations.

Proposition 3. Let \mathbf{X} be a statistical vector and Z a concomitant variable, then covariance matrix $\Sigma_{\mathbf{X}}$ can be decomposed into three terms,

$$\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X}}^* + {}_L\Sigma_{\mathbf{X}} + {}_C\Sigma_{\mathbf{X}}. \quad (34)$$

Proposition 4. Let \mathbf{X} be a statistical vector and Z a concomitant variable. *Partial covariance* matrix of \mathbf{X} after removing the linear effects of Z , ${}_Z\Sigma_{\mathbf{X}} = E[(\mathbf{X} - \hat{\mathbf{A}}\mathbf{Z})(\mathbf{X} - \hat{\mathbf{A}}\mathbf{Z})']$, is the sum of the first two components of Equation (34), i.e., residual covariance matrix and covariance lack of fit matrix,

$${}_Z\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X}}^* + {}_L\Sigma_{\mathbf{X}}. \quad (35)$$

Proof.

$$\begin{aligned}
{}_Z\Sigma_{\mathbf{X}} &= E[(\mathbf{X} - \hat{\mathbf{A}}\mathbf{Z})(\mathbf{X} - \hat{\mathbf{A}}\mathbf{Z})'] \\
&= E[(\mathbf{X} \pm \mu_{\mathbf{X}}(Z) - \hat{\mathbf{A}}\mathbf{Z})(\mathbf{X} \pm \mu_{\mathbf{X}}(Z) - \hat{\mathbf{A}}\mathbf{Z})'] \\
&= E_Z(\Sigma_{\mathbf{X}}(Z)) + E_Z[(\mu_{\mathbf{X}}(Z) - \hat{\mathbf{A}}\mathbf{Z})(\mu_{\mathbf{X}}(Z) - \hat{\mathbf{A}}\mathbf{Z})'] \\
&= \Sigma_{\mathbf{X}}^* + {}_L\Sigma_{\mathbf{X}}.
\end{aligned}$$

10 Concluding remarks

The proposed decompositions of covariance in Equations (10) and (34) in univariate and multivariate cases, constitute a fruitful basis for comparative evaluation of correlation at different levels, both individual and aggregate.

Some paradoxical aspects of aggregate analysis in linear model building may be adequately resolved by recognizing that the choice of a special reference population in statistics is a central issue that determines, to some extent, analysis and interpretation of results.

Reconstruction of individual level relationships on the basis of aggregate level frameworks is quite limited and depends upon further assumptions that may modify the original problem and related conclusions.

References

- Achen CH, Shilvelly WP (1995) *Cross-Level Inference*. University of Chicago Press, Chicago
- Faraway JJ (2005) *Linear Models with R*. Chapman and Hall/CRC Press, Boca Raton
- Freedman DA, Klein SP, Sacks J, Smyth CA, Everett CG (1991) Ecological regression and voting rights. *Evaluation Review* 15: 673–711 (with discussion)
- Freedman DA, Klein SP, Ostland M, Roberts MR (1998) Review of A Solution to the Ecological Inference Problem. *Journal of The American Statistical Association* 93: 1518–1522 with discussion, vol. 94 (1999), 352–357
- Guseo R (2006) *Statistica*, 3rd edition. Cedam, Padua
- King G (1997) *A Solution to the Ecological Inference Problem*. Princeton University Press, Princeton
- King G, Rosen O, Tanner MA (2004) *Ecological Inference: new methodological strategies*. Cambridge University Press, Cambridge
- Robinson WS (1950) Ecological correlations and the behavior of individuals. *American Sociological Review* 15: 351–357
- Schuessler AA (1999) Ecological Inference. *Proceedings of the National Academy of Sciences USA* 96: 10578–10581