

Partial and Ecological Correlation: a Common Three Terms Covariance Decomposition ¹

Correlazione Parziale ed Ecologica: una Comune Scomposizione a Tre Termini della Covarianza

Renato Guseo ²
Dipartimento di Scienze Statistiche,
Università degli Studi di Padova, e-mail: renato.guseo@unipd.it

Riassunto: La covarianza tra le variabili X e Y in presenza di una variabile concomitante Z , viene scomposta in tre termini. La covarianza parziale è definita dalla somma dei primi due. La covarianza ecologica è costituita dalla somma degli ultimi due. Tale scomposizione consente un'analisi comparata della *ecological fallacy* rispetto alle procedure di selezione step-wise rette dalla correlazione parziale. E' ben noto il rischio di una previsione individuale grossolanamente errata basata su dati aggregati. Un ulteriore esempio attiene allo studio di un modello nidificato a due livelli. La separazione tra il caso descrittivo ed il caso stocastico è inessenziale.

Keywords: partial correlation, ecological correlation, ecological fallacy

1. Introduction

Let us consider, for simplicity sake, a bivariate statistical variable (X, Y) . The well-known two terms variance decomposition highlights the joint role of conditional variances, $\sigma_Y^2(x)$, $x \in X$, and regression function, $\mu_Y(x)$, $x \in X$, i.e., $\sigma_Y^2 = M_X(\sigma_Y^2(X)) + \bar{\sigma}_Y^2 = \sigma_Y^{*2} + \bar{\sigma}_Y^2$, where $\bar{\sigma}_Y^2 = M_X[(\mu_Y(X) - \mu_Y)^2]$ is the *explained variance* and σ_Y^{*2} denotes the *residual variance*. Previous decomposition is a special case of more general covariance decompositions, Guseo (2006). *Ecological fallacy* is a possible risk in linear models prediction 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 referred to Freedman et al. (1991), Achen and Shilvely (1995), King (1997), Freedman et al. (1998) and Schuessler (1999). If we study only individual data such a paradox does not exist.

2. Covariance decompositions

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

$$\sigma_{XY} = \sigma_{XY}^* + \sigma_{\bar{X}\bar{Y}}. \quad (1)$$

¹Financial support was provided by Miur. See <http://homes.stat.unipd.it/guseo/> for an extended version.

²Address of correspondence: via C. Battisti 241, 35100 Padova, Italy.

Usual *explained variance* is conceptually equivalent to *ecological variance*. In fact, for $X = Y$, we have $\sigma_{YY} = \sigma_Y^2 = \sigma_Y^{*2} + \sigma_Y^2 = \sigma_Y^{*2} + \bar{\sigma}_Y^2$. Ecological covariance represents the direct covariance between regression functions $\mu_X(Z)$ and $\mu_Y(Z)$. We may study such a relationship by introducing a deviation from the corresponding linear models determined under the least squares method, e.g., $X = a + bZ + \varepsilon_1$ and $Y = c + dZ + \varepsilon_2$.

Proposition 2. Let (X, Y) be a bivariate statistical variable and Z a concomitant variable. 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*, ${}_C\sigma_{XY}$, where ${}_L\sigma_{XY} = M_Z(\mu_X(Z) - \hat{a} - \hat{b}Z)(\mu_Y(Z) - \hat{c} - \hat{d}Z)$ and ${}_C\sigma_{XY} = M_Z(\hat{a} + \hat{b}Z - \mu_X)(\hat{c} + \hat{d}Z - \mu_Y) = (\sigma_{XZ}\sigma_{YZ})/\sigma_Z^2$,

$$\sigma_{\bar{X}\bar{Y}} = {}_L\sigma_{XY} + {}_C\sigma_{XY}. \quad (2)$$

Proposition 3. Let (X, Y) be a bivariate statistical variable and Z a concomitant variable. A three terms decomposition of covariance is

$$\sigma_{XY} = \sigma_{XY}^* + {}_L\sigma_{XY} + {}_C\sigma_{XY}, \quad (3)$$

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

Proposition 4. Let (X, Y) be a bivariate statistical variable and Z a concomitant variable. *Partial covariance* between X and Y under the exclusion of linear effects of Z , $\sigma_Z(X, Y) = M[(X - \hat{a} - \hat{b}Z)(Y - \hat{c} - \hat{d}Z)]$, is the sum of the first two components of Equation (3), i.e., residual covariance and covariance lack of fit,

$$\sigma_Z(X, Y) = \sigma_{XY}^* + {}_L\sigma_{XY}. \quad (4)$$

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

Equations (1–4) allow a comparative assessment of ecological fallacy and step-wise model building in regressive nonparametric contexts based on partial correlation. The latter avoids risky sign errors in linear models component effects inferences that affect predictions. Further applications refer to hierarchical model analysis.

References

- Achen C.H., Shilvely W.P. (1995) *Cross-Level Inference*, University of Chicago Press, Chicago.
- Freedman D.A., Klein S.P., Sacks J., Smyth C.A., Everett C.G. (1991) Ecological regression and voting rights, *Evaluation Review*, 15, 673–711 (with discussion).
- Freedman D.A., Klein S.P., Ostland M., Roberts M.R. (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*, 3-rd Edition, Cedam, Padova.
- King G. (1997) *A Solution to the Ecological Inference Problem*, Princeton University Press, Princeton.
- Robinson W.S. (1950) Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357.
- Schuessler A.A. (1999) Ecological Inference. *Proceedings of the National Academy of Sciences USA*, 96, 10578–10581.