

Outliers in hierarchical models: an application to air pollution data

Osservazioni anomale nei modelli gerarchici: un'applicazione ai dati di inquinamento atmosferico

Miriam Daniele

Dipartimento di Statistica e Matematica "S. Vianelli",
Università di Palermo, e-mail: mdaniele@dssm.unipa.it

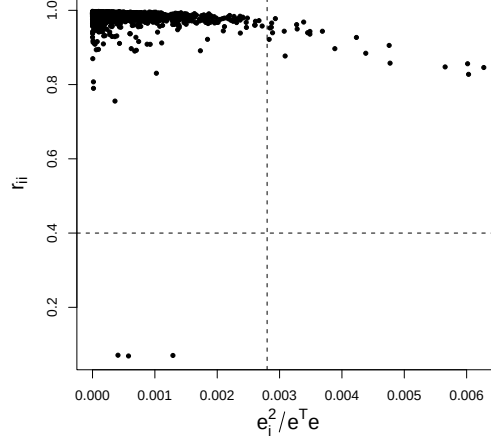
Riassunto: L'analisi delle osservazioni anomale rappresenta una valida tecnica esplorativa per lo studio di tutti quei fenomeni reali in cui queste osservazioni sono quelle di primario interesse. Nell'ambito del monitoraggio dell'inquinamento atmosferico, ad esempio, le osservazioni anomale spesso sono quei valori che superano considerevolmente i valori soglia fissati dalla normativa vigente. Nel presente lavoro, dopo aver studiato l'effetto di alcune variabili meteo sul PM_{10} , attraverso un modello lineare gerarchico, si indaga sulla presenza di osservazioni anomale. Seguendo l'impostazione più diffusa in letteratura si utilizzano procedure informali basate sui residui studentizzati e sui valori che hanno un maggior effetto leva.

Keywords: Outliers, HLM, PM_{10} , weather variables

1. Introduction

Hierarchical models are useful tool in describing relationships between a response variable and covariates in data that are grouped according to some classification factors. These models encompass two sources of variation, namely within and among individuals in the population; this naturally leads to the possibility of contaminated data at each sampling level (Yeap and Davidian (2001)). We refer to such data as outliers when they appear inconsistent with the assumed model for the rest of the data. Outlier analysis is enough complex in a multilevel dataset, because the researcher has to consider at what levels a particular response is outlying; furthermore, the treatment of a particular response at one level may affect its status or the status of other units at other levels in the model (Langford and Lewis, 1998). Formal and informal procedures of diagnostic, such as residuals analysis and detection of leverage points and outliers, are a little explored area of hierarchical modelling, particularly in a nonlinear context. In this paper we focus our analysis on data of air pollution collected in 8 monitoring sites in Palermo in 2003. Due to the hierarchy of the dataset, we fitted a hierarchical linear model to study the effects of some weather variables on daily average concentrations of PM_{10} (logarithmic transformation). In the final model the significative variables result to be the weather variables air temperature, wind direction and wind speed; in particular, the fixed part involves the air temperature and the wind direction, the random part the air temperature, the wind direction and the wind speed. Aim of this paper is to detect outliers and leverage, via a residual analysis of fitted model, in order to inspect about the observations largely exceeding threshold values and to monitor the risk of very dangerous exposures.

Figure 1: Scatter plot of r_{ii} versus $e_i^2/e^T e$



2. Application

The general form of the two level hierarchical linear model fitted to the data is

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\beta}_{(p \times 1)}$ is a vector of unknown constants, the fixed effects of the model; $X_{(n \times p)}$ is a design matrix of fixed numbers associated with $\boldsymbol{\beta}$; $Z = [Z_1 | \dots | Z_r]$, where Z_i is a $(n \times q_i)$ known design matrix of the random effect factor i ; $\mathbf{u}' = [\mathbf{u}'_1 | \dots | \mathbf{u}'_r]$, where $\mathbf{u}'_i$ is a $(q_i \times 1)$ vector of random variables from $N(\mathbf{0}, \sigma_i^2 I)$, $i = 1, \dots, r$; $\boldsymbol{\varepsilon}_{(n \times 1)}$ is an vector of error terms from $N(\mathbf{0}, \sigma_e^2 I)$, with $\mathbf{u}'_i$ and $\boldsymbol{\varepsilon}$ mutually independent. One may also write $\mathbf{u} \sim N(\mathbf{0}, \sigma_e^2 D)$, where D is block diagonal with the i -th block being $\gamma_i I_{q_i}$, for $\gamma_i = \sigma_i^2 / \sigma_e^2$, so that $\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma_e^2 H)$, for $H = I + \sum_{i=1}^r \gamma_i Z_i Z_i'$. The fitted value of the response variable $\mathbf{y}$ are $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} + Z\tilde{\mathbf{u}}$, giving the residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 R)$, where $R = H^{-1} - H^{-1}X(X'VX)^{-1}X'H^{-1}$ ($V = \sigma_e^2 I$).

Fitted the model by the REML method and computed the R matrix we can plot (shown in Figure 1) its diagonal elements versus a transformation of residuals at Level-1 e_i (as suggested by Zewotir and Galpin (2006)), that is effective in displaying the outliers and leverages simultaneously. Points are expected to concentrate around the upper-left corner of the plot. Points that are separated from the bulk of the other points and fall in the lower-left corner of the scatter plot are high-leverage points and points that are separated to the right side are outliers, with high leverage those in the lower-right corner.

We are still analyzing these outliers or high-leverage points in order to assess their influence especially on the model estimation.

References

- Langford I.H. and Lewis T. (1998) Outliers in multilevel data, *J. R. Stat. A*, 161, 121–160.
 Yeap B.Y. and Davidian M. (2001) Robust Two-Stage Estimation in Hierarchical Nonlinear Models, *Biometrics*, 57, 266–272.
 Zewotir T. and Galpin J.S. (2006) A unified approach on Residuals, Leverages and Outliers in the Linear Mixed Model, *Test 16. In Press*.