

Influence analysis in small area estimation

Morales D., Pagliarella M.C. and Salvatore R.

Abstract This paper discusses the leverage matrix for the Fay-Herriot area-level model, when the area-random effect variance is estimated by the restricted maximum likelihood method. Leverage and Cook's distance are some of the most important tools in influence analysis, where the main target is to identify observations or groups of observations that might determine the character of model estimates and predictors. In the small area estimation setup, applied statistician are interested in tools to identify observation that might influence the variance component and the regression parameter estimates, the EBLUP and its mean squared error estimate. This paper partially address these issues.

Key words: small area estimation, influence analysis, leverage, Cook's distance.

1 Introduction

Two principal types of influence analysis for linear models are developed. In the first, the calculation of leverage (defined as the diagonal elements of the hat matrix) plays a central role. The second is based on the influence of case deletion and is

Domingo Morales
Universidad Miguel Hernández de Elche, Spain, e-mail: d.morales@umh.es

Maria Chiara Pagliarella
Università degli studi di Cassino, Italy, e-mail: mc.pagliarella@unicas.it

Renato Salvatore
Università degli studi di Cassino, Italy, e-mail: rsalvatore@unicas.it

This paper presents some of the results from the PRIN 2007 Italian National Research Project: "The use of auxiliary information in complex sampling survey designs", granted by the Italian Ministry of University and Education. The authors thank the Italian National Statistics Institute for supplying the data.

called case deletion diagnostics, i.e. case deletion diagnostic approach suggested by Cook (1977). The concept of influence is understood as sensitivity of a model studied through measures that express its stability under perturbations. Changes in the data or model components might produce significant changes in the model output. So, the goal of influence analysis is to determine which cases are influential and the manner in which they are important to the analysis.

This paper discusses the leverage matrix for the Fay-Herriot area-level model, when we deal either with restricted maximum likelihood (REML) or with moment estimates of the random-area effect variance. Section 2 introduces the area-level Fay-Herriot model, derives an approximate closed-form solution for the REML random-area effect variance estimator. The obtained expression of REML estimator is useful for detecting influential observations (the area direct estimates) that affect the estimate model variance. Section 3 gives the leverage matrix on the fixed effects (the marginal leverage matrix) and the leverage matrix on random-area effects (the random effect leverage matrix, see Demidenko (2005), Fung et al. (2002), and Wey et al. (1998)). Section 4 shows the influence analysis on the first two terms of the estimated mean squared error of the EBLUP. Section 5 presents some Cook's distances. Section 6 introduce an application to real data. Finally Section 7 gives some conclusions.

2 The Fay-Herriot model

Consider the traditional Fay-Herriot model, in which the $\hat{y}_i$'s are the area direct estimates, or functions of it.

$$\hat{y}_i = \mathbf{x}_i' \boldsymbol{\beta} + b_i v_i + e_i, \quad e_i \stackrel{ind}{\sim} (0, \psi_i), \quad v_i \stackrel{iid}{\sim} (0, \sigma_v^2), \quad i = 1, \dots, m.$$

We rewrite the model in the general matrix form

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}^{1/2}\mathbf{v} + \mathbf{e}, \quad \mathbf{B} = \text{diag}(b_i^2), \quad \text{var}(\mathbf{y}) = \mathbf{V} = \text{diag}(V_i) = \text{diag}(\psi_i + \sigma_v^2 b_i^2).$$

The vector of the BLUP's is as follows,

$$\hat{\mathbf{y}}^H = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{B}^{1/2}\hat{\mathbf{v}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{B}^{1/2}\sigma_v^2\mathbf{B}^{1/2}\mathbf{V}^{-1}(\hat{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}\hat{\boldsymbol{\beta}} + \Gamma(\hat{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

with $\Gamma = \text{diag}(\gamma_i) = \text{diag}(\sigma_v^2 b_i^2 / (\psi_i + \sigma_v^2 b_i^2))$. The generalized least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\hat{\mathbf{y}}),$$

By using the relation $\mathbf{V}^{-1} = (\boldsymbol{\Psi} + \sigma_v^2\mathbf{B})^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi}^{-1} + (\sigma_v^2)^{-1}\mathbf{B}^{-1})^{-1}\boldsymbol{\Psi}^{-1} = \boldsymbol{\Psi}^{-1}(\mathbf{I} - \Gamma)$, with $\boldsymbol{\Psi} = \text{diag}(\psi_i)$ and $\mathbf{X}^* = \boldsymbol{\Psi}^{-1/2}\mathbf{X}$, one gets

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GLS} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\hat{\mathbf{y}}) = (\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X} - \mathbf{X}'\boldsymbol{\Psi}^{-1}\Gamma\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Psi}^{-1}\hat{\mathbf{y}} - \mathbf{X}'\boldsymbol{\Psi}^{-1}\Gamma\hat{\mathbf{y}}) \\ &= (\mathbf{X}^{*'}\mathbf{X}^* - \mathbf{X}^{*'}\Gamma\mathbf{X}^*)^{-1}(\mathbf{X}^{*'}\hat{\mathbf{y}}^* - \mathbf{X}^{*'}\Gamma\hat{\mathbf{y}}^*). \end{aligned}$$

If $\Gamma = \mathbf{0}$, the $\hat{\beta}_{GLS}$ becomes the weighted least squares estimator $\hat{\beta}_{WLS}$, with weight matrix $\Psi^{-1/2}$. When both $\Gamma = \mathbf{0}$ and $\psi_i = \psi$ among the areas, $\hat{\beta}_{WLS}$ reduces to the ordinary least squares estimator $\hat{\beta}_{OLS}$. Let us now define

$$\Gamma^* = \text{diag}(\gamma^*) = \text{diag}(\sigma_v^2/(\phi^* + \sigma_v^2)),$$

where $\gamma^* \text{vec}'(\mathbf{X}^*) \text{vec}(\mathbf{X}^*) = \text{vec}'(\mathbf{X}^*)(\mathbf{I}_p \otimes \Gamma) \text{vec}(\mathbf{X}^*)$. It holds that $\gamma^* = c/a$, $c = \text{vec}'(\mathbf{X}^*)(\mathbf{I}_p \otimes \Gamma) \text{vec}(\mathbf{X}^*)$, $a = \text{vec}'(\mathbf{X}^*) \text{vec}(\mathbf{X}^*)$, where the symbol $\otimes$ denotes the Kronecker product, and vec the operator that vectorizes $\mathbf{X}^*$ by stacking its columns. The ratios $\gamma^* = c/a$, $\sigma_v^2/(\phi^* + \sigma_v^2) = c/a$ and $\phi^* = (a/c - 1)\sigma_v^2$ can be calculated by using the formulas $V_i = \psi_i + \sigma_v^2 b_i^2 \approx \phi^* b_i^2 + \sigma_v^2 b_i^2 = (a/c)\sigma_v^2 b_i^2$, $V_i^{-1} \approx c/(a\sigma_v^2 b_i^2)$, $V_i^{-1} = 1/\psi_i - \gamma_i/\psi_i \approx 1/\psi_i - (1/\psi_i)c/a = (a-c)/\psi_i a$, $V_i = \psi_i a/(a-c)$. Note that c is again function of σ_v^2 , but the values of a and c can be computed beforehand. If we start now from the restricted version of the likelihood function,

$$\begin{aligned} l_{REML}(\beta, \sigma_v^2) = & -\frac{m-p}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}^{*'} \mathbf{V}^{-1} \mathbf{X}^*| \\ & - \frac{1}{2} (\hat{\mathbf{y}}^* - \mathbf{X}^* \hat{\beta}_{WLS})' \mathbf{V}^{-1} (\hat{\mathbf{y}}^* - \mathbf{X}^* \hat{\beta}_{WLS}), \end{aligned}$$

and we rewrite the $\hat{\beta}_{GLS}$ estimator in the form

$$\hat{\beta}_{GLS} = (\mathbf{X}^{*'} \mathbf{X}^* - \gamma^* \mathbf{X}^{*'} \mathbf{X}^*)^{-1} (\mathbf{X}^{*'} \hat{\mathbf{y}}^* - \gamma^* \mathbf{X}^{*'} \hat{\mathbf{y}}^*),$$

we can easily show that $\hat{\beta}_{GLS} \equiv \hat{\beta}_{WLS}$, while REML function takes the form

$$\begin{aligned} l_{REML}(\gamma^*) = & -\frac{1}{2} \left[(m-p) \log 2\pi - (m-1) \log(1-\gamma^*) + \sum \left\| \hat{\mathbf{y}}^* - \mathbf{X}^* \hat{\beta}_{WLS} \right\|^2 \right. \\ & \left. - \gamma^* \sum (\hat{y}_i^* - \bar{y}^*)^2 \right]. \end{aligned}$$

Since $\hat{\beta}_{GLS} = \hat{\beta}_{WLS}$ and independent of $\mathbf{V}$, by taking derivatives on l_{REML} with respect to γ^* , we obtain the closed-form solution for γ^* , i.e.

$$\frac{\partial l_{REML}(\gamma^*)}{\partial (\gamma^*)} = -\frac{1}{2} \left[\frac{m-1}{1-\gamma^*} - \sum (\hat{y}_i^* - \bar{y}^*)^2 \right] = 0, \quad \hat{\gamma}_{REML} = \frac{\sum (\hat{y}_i^* - \bar{y}^*)^2 - (m-1)}{\sum (\hat{y}_i^* - \bar{y}^*)^2}.$$

Applying the first-order approximation $\sigma_v^2 b_i^2 / (\psi_i + \sigma_v^2 b_i^2) \approx \psi_i^{-1} \sigma_v^2 b_i^2$, we get

$$c = \text{vec}'(\mathbf{X}^*)(\mathbf{I}_p \otimes \Gamma) \text{vec}(\mathbf{X}^*) \approx \sigma_v^2 \text{vec}'(\mathbf{X}^*)(\mathbf{I}_p \otimes \Phi) \text{vec}(\mathbf{X}^*) = \sigma_v^2 c^*,$$

where $\Phi = \text{diag}(\phi_i^{-1})$, $\phi_i = \psi_i/b_i^2$ and $\sigma_v^2 b_i^2 / (\psi_i + \sigma_v^2 b_i^2) \approx \psi_i^{-1} \sigma_v^2 b_i^2$. Substituting the corresponding values on the estimate of $\hat{\beta}_{WLS}$, we obtain

$$\gamma^* = \frac{\sigma_v^2}{\phi^* + \sigma_v^2} = \frac{c}{a} \approx \frac{\sigma_v^2 c^*}{a}, \quad \hat{\sigma}_{v,REML}^2 = \frac{a}{c^*} \frac{\sum (\hat{y}_i^* - \bar{\hat{y}}^*)^2 - (m-1)}{\sum (\hat{y}_i^* - \bar{\hat{y}}^*)^2}.$$

3 The leverage matrix

Since the leverage matrix for the traditional mixed model is $\frac{\partial \hat{\mathbf{y}}^H}{\partial \hat{\mathbf{y}}}$, in the sequel we compute some useful vector and matrix derivatives, in order to get the leverage matrix of fixed and random effects. We start from the definition of A as $\sum (\hat{y}_i^* - \bar{\hat{y}}^*)^2 = \hat{\mathbf{y}}^{*\prime} \hat{\mathbf{y}}^* - \frac{1}{m} (\hat{\mathbf{y}}^{*\prime} \mathbf{1}_m)^2$. We have the following relations:

- 1) $\frac{\partial A}{\partial \hat{\mathbf{y}}} = \frac{\partial}{\partial \hat{\mathbf{y}}} \sum (\hat{y}_i^* - \bar{\hat{y}}^*)^2 = \frac{\partial}{\partial \hat{\mathbf{y}}} [\hat{\mathbf{y}}^{*\prime} \hat{\mathbf{y}}^* - \frac{1}{m} (\hat{\mathbf{y}}^{*\prime} \mathbf{1}_m)^2] = 2\hat{\mathbf{y}}' \Psi^{-1} - \frac{2}{m} (\hat{\mathbf{y}}' \mathbf{1}_m) (\mathbf{1}_m')',$
 $\mathbf{1}_m' = \Psi^{-1/2} \mathbf{1}_m$
- 2) $\frac{\partial \hat{\sigma}_{v,REML}^2}{\partial \hat{\mathbf{y}}} = \frac{\partial}{\partial \hat{\mathbf{y}}} \left[\frac{a}{c^*} \frac{A - (m-1)}{A} \right] = \frac{a}{c^*} A^{-2} [2\hat{\mathbf{y}}' \Psi^{-1} - \frac{2}{m} (\hat{\mathbf{y}}' \mathbf{1}_m) (\mathbf{1}_m')'] = \sigma_\delta^2$
- 3) $\frac{\partial \hat{\mathbf{V}}}{\partial \hat{\mathbf{y}}} = \frac{\partial}{\partial \hat{\mathbf{y}}} \text{diag}(\psi_i + \hat{\sigma}_{v,REML}^2 b_i^2) = \frac{\partial}{\partial \hat{\mathbf{y}}} (\Psi + \mathbf{B}^{1/2} \hat{\sigma}_{v,REML}^2 \mathbf{B}^{1/2}) = \mathbf{B} \otimes \sigma_\delta^2$
- 4) $\frac{\partial \hat{\mathbf{V}}^{-1}}{\partial \hat{\mathbf{y}}} = -(\hat{\mathbf{V}}^{-1} \mathbf{B} \hat{\mathbf{V}}^{-1}) \otimes \sigma_\delta^2$
- 5) $\frac{\partial (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}}{\partial \hat{\mathbf{y}}} = -\frac{\partial (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})}{\partial \hat{\mathbf{y}}} [(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X}) \otimes (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})]$
 $= [(\hat{\mathbf{V}}^{-1} \mathbf{B} \hat{\mathbf{V}}^{-1}) \otimes \sigma_\delta^2] (\mathbf{X} \otimes \mathbf{X}) [(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X}) \otimes (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})].$

Taking into account that the leverage matrix with fixed $\mathbf{V}$ is

$$L(\hat{\beta}, \hat{\mathbf{v}}) = L(\hat{\beta}) + L(\hat{\mathbf{v}}) = \mathbf{H}_1 + \mathbf{H}_2 = \mathbf{X}(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1} + \hat{\sigma}_{v,REML}^2 \mathbf{B} \hat{\mathbf{P}},$$

with $\mathbf{H}_1 = \mathbf{X}(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1}$, $\mathbf{H}_2 = \hat{\sigma}_{v,REML}^2 \mathbf{B} \hat{\mathbf{V}}^{-1} (\mathbf{I}_m - \mathbf{H}_1)$ and $\hat{\mathbf{P}} = \hat{\mathbf{V}}^{-1} - \hat{\mathbf{V}}^{-1} \mathbf{X}(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1}$. The explicit form, in terms of sampled observations, of the leverage matrix is

$$L^*(\hat{\beta}, \hat{\mathbf{v}}) = \frac{\partial \hat{\mathbf{y}}^H}{\partial \hat{\mathbf{y}}} = \frac{\partial (\mathbf{X} \hat{\beta})}{\partial \hat{\mathbf{y}}} + \frac{\partial (\mathbf{B}^{1/2} \hat{\mathbf{v}})}{\partial \hat{\mathbf{y}}} = L^*(\hat{\beta}) + L^*(\hat{\mathbf{v}}).$$

Based on the derivative

$$\frac{\partial \mathbf{H}_1}{\partial \hat{\mathbf{y}}} = \left[\frac{\partial (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}}{\partial \hat{\mathbf{y}}} (\mathbf{X}' \otimes \mathbf{X}') \right] (\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_m) + \frac{\partial \hat{\mathbf{V}}^{-1}}{\partial \hat{\mathbf{y}}} (\mathbf{I}_m \otimes [\mathbf{X}(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}']),$$

the leverage matrix for the fixed effects is given by

$$L^*(\hat{\beta}) = \frac{\partial (\mathbf{X} \hat{\beta})}{\partial \hat{\mathbf{y}}} = \frac{\partial}{\partial \hat{\mathbf{y}}} [\mathbf{X}(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1} \hat{\mathbf{y}}] = \left(\frac{\partial \mathbf{H}_1}{\partial \hat{\mathbf{y}}} \right) (\hat{\mathbf{y}} \otimes \mathbf{I}_m) + \mathbf{H}_1' = \mathbf{H}_1^* + \mathbf{H}_1'.$$

Denoting with $\mathbf{H}_{1,ii}^*$ the diagonal element of the matrix $\mathbf{H}_1^*$ for the i -th area, and remembering that $\text{tr}(\mathbf{H}_1') = p$, influential observations that affect the fixed-effects es-

timates, through the estimation of the model variance more directly, can be checked comparing the quantity $(\mathbf{H}_{1,ii}^* - \frac{1}{m}\text{tr}(\mathbf{H}_1^*))$ with $[L^*(\hat{\beta})]_{ii} - \frac{1}{m}\text{tr}(L^*(\hat{\beta}))$. When we assess the potentially influential values, this is very useful in analyzing the contribution to the leverage of the single observation in estimating the model variance, respect to the marginal leverage. The leverage for the random effects is

$$\begin{aligned} L^*(\hat{\mathbf{v}}) &= \frac{\partial(\mathbf{B}^{1/2}\hat{\mathbf{v}})}{\partial\hat{\mathbf{y}}} = \frac{\partial}{\partial\hat{\mathbf{y}}}[\mathbf{B}^{1/2}\hat{\sigma}_{v,REML}^2\mathbf{B}^{1/2}\hat{\mathbf{V}}^{-1}(\hat{\mathbf{y}} - \mathbf{X}\hat{\beta})] \\ &= [(\mathbf{B} \otimes \sigma_\delta^2)(\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_m) + (\frac{\partial\hat{\mathbf{V}}^{-1}}{\partial\hat{\mathbf{y}}})(\mathbf{I}_m \otimes \hat{\sigma}_{v,REML}^2\mathbf{B})][(\hat{\mathbf{y}} - \mathbf{X}\hat{\beta}) \otimes \mathbf{I}_m] \\ &\quad + [\mathbf{I}_m - L^*(\hat{\beta})](\hat{\sigma}_{v,REML}^2\mathbf{B}\hat{\mathbf{V}}^{-1}). \end{aligned}$$

Combining the above results, we get

$$\begin{aligned} L^*(\hat{\beta}, \hat{\mathbf{v}}) &= L^*(\hat{\beta})(\mathbf{I}_m - \hat{\sigma}_{v,REML}^2\mathbf{B}\hat{\mathbf{V}}^{-1}) + (\hat{\sigma}_{v,REML}^2\mathbf{B}\hat{\mathbf{V}}^{-1}) \\ &\quad + [(\mathbf{B} \otimes \sigma_\delta^2)(\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_m) + (\frac{\partial\hat{\mathbf{V}}^{-1}}{\partial\hat{\mathbf{y}}})(\mathbf{I}_m \otimes \hat{\sigma}_{v,REML}^2\mathbf{B})][(\hat{\mathbf{y}} - \mathbf{X}\hat{\beta}) \otimes \mathbf{I}_m] \\ &= L^*(\hat{\beta})[\mathbf{I}_m - \mathbf{H}_2(\mathbf{I}_m - \mathbf{H}_1)^{-1}] + \mathbf{H}_2(\mathbf{I}_m - \mathbf{H}_1)^{-1} + \mathbf{T} \\ \mathbf{T} &= [(\mathbf{B} \otimes \sigma_\delta^2)(\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_m) + (\frac{\partial\hat{\mathbf{V}}^{-1}}{\partial\hat{\mathbf{y}}})(\mathbf{I}_m \otimes \hat{\sigma}_{v,REML}^2\mathbf{B})][(\hat{\mathbf{y}} - \mathbf{X}\hat{\beta}) \otimes \mathbf{I}_m]. \end{aligned}$$

4 Influence analysis on the MSE area estimates

Now take a look on the influence of observations on the estimation of the mean squared error (MSE) of the EBLUP estimates. With regard to component $\mathbf{G}_1 = \text{diag}(g_{1i}) = \Gamma\Psi$ of MSE, influential area estimates can be detected by

$$\begin{aligned} \frac{\partial\mathbf{G}_1}{\partial\hat{\mathbf{y}}} &= \frac{\partial}{\partial\hat{\mathbf{y}}}(\Gamma\Psi) = \frac{\partial}{\partial\hat{\mathbf{y}}}(\hat{\sigma}_{v,REML}^2\mathbf{B}\hat{\mathbf{V}}^{-1}\Psi) \\ &= (\mathbf{B} \otimes \sigma_\delta^2)(\hat{\mathbf{V}}^{-1}\Psi \otimes \mathbf{I}_m) + [\frac{\partial\hat{\mathbf{V}}^{-1}}{\partial\hat{\mathbf{y}}}(\Psi \otimes \mathbf{I}_m)](\mathbf{I}_m \otimes \hat{\sigma}_{v,REML}^2\mathbf{B}). \end{aligned}$$

Finally we have the influence on the mean squared error component related to the variation of the fixed effects. Starting from the matrix

$$\mathbf{G}_2 = (\mathbf{I}_m - \Gamma)\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}_m - \Gamma) = (\mathbf{I}_m - \Gamma)\mathbf{U}(\mathbf{I}_m - \Gamma),$$

we have

$$\frac{\partial\mathbf{G}_2}{\partial\hat{\mathbf{y}}} = \frac{\partial(\mathbf{I}_m - \Gamma)}{\partial\hat{\mathbf{y}}}([\mathbf{U}(\mathbf{I}_m - \Gamma)] \otimes \mathbf{I}_m) + \frac{\partial[\mathbf{U}(\mathbf{I}_m - \Gamma)]}{\partial\hat{\mathbf{y}}}[\mathbf{I}_m \otimes (\mathbf{I}_m - \Gamma)],$$

where

$$\begin{aligned}\frac{\partial(\mathbf{I}_m - \Gamma)}{\partial \hat{\mathbf{y}}} &= -(\mathbf{B} \otimes \sigma_\delta^2)(\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_m) + \left(\frac{\partial \hat{\mathbf{V}}^{-1}}{\partial \hat{\mathbf{y}}} \mathbf{I}_{m^2}\right)(\mathbf{I}_m \otimes \hat{\sigma}_{v,REML}^2 \mathbf{B}), \\ \frac{\partial[\mathbf{U}(\mathbf{I}_m - \Gamma)]}{\partial \hat{\mathbf{y}}} &= \left[\frac{\partial(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}}{\partial \hat{\mathbf{y}}}(\mathbf{X}' \otimes \mathbf{X}')\right][(\mathbf{I}_m - \Gamma) \otimes \mathbf{I}_m] + \frac{\partial(\mathbf{I}_m - \Gamma)}{\partial \hat{\mathbf{y}}}(\mathbf{I}_m \otimes \mathbf{U}).\end{aligned}$$

5 Cluster deletion diagnostics and Cook's distance

Cook's distance assesses the effects of a more drastic (global) change by removing an entire data point or clusters. In this section subindex (ℓ) is used for estimators that are calculated after deleting case ℓ . The Cook's distance for $\hat{\sigma}_v^2$ is

$$d_\ell^v = \frac{(\hat{\sigma}_v^2 - \hat{\sigma}_{v(\ell)}^2)^2}{\widehat{\text{var}}(\hat{\sigma}_v^2)},$$

where $\text{var}(\hat{\sigma}_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2$ that is obtained from the inverse of the REML Fisher information matrix. The proposed Cook-type distance for the EBLUP $\hat{y}_i^H$ is

$$d_\ell^{eblup} = \frac{(\hat{y}_i^H - \hat{y}_{i\ell}^H)^2}{\text{mse}(\hat{y}_i^H)},$$

where

$$\text{mse}(\hat{y}_i^H) = g_1(\hat{\sigma}_u^2) + g_2(\hat{\sigma}_u^2) + 2g_3(\hat{\sigma}_u^2)$$

is the Prasad-Rao (1990) estimator of the mean squared error of $\hat{y}_i^H$, with

$$\begin{aligned}g_1(\sigma_v^2) &= \frac{\sigma_v^2 \sigma_i^2}{\sigma_v^2 + \sigma_i^2}, \quad g_2(\sigma_v^2) = \frac{\sigma_i^4}{(\sigma_v^2 + \sigma_i^2)^2} \mathbf{x}_i(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i', \\ g_3(\sigma_v^2) &= \frac{\sigma_i^4}{(\sigma_v^2 + \sigma_i^2)^3} \text{var}(\hat{\sigma}_v^2).\end{aligned}$$

6 An Application

In this section we present an application of the Fay-Harriot area-level model to official Agricultural Census Data (Survey structure and production of Italian farms (SPA) year 2007), with the aim to compare the estimates of the marginal leverage matrix of the fixed effects as developed in Section 3. The dataset contains 103 observations (areas in our context) that are the Italian provinces. The target variable is the mean of standard gross profit. The selected auxiliary variables are the mean of irrigable area and the number of worked days. We estimate the REML random-area

effect variance and we calculate the derivatives for the construction of the leverage matrix of the fixed effects in two ways: $L(\hat{\beta})$ and $L^*(\hat{\beta})$. We present the results in the scatter plots appearing in Figures 1 and Figure 2.

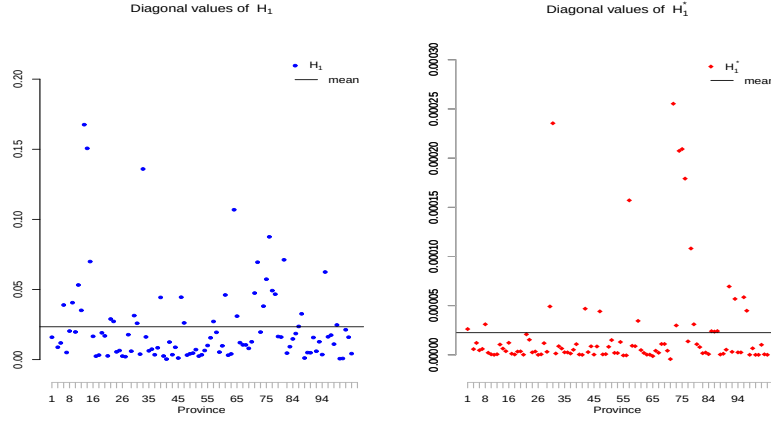


Fig. 1 Comparison between the diagonal values of matrices $\mathbf{H}_1$ and $\mathbf{H}_1^*$.

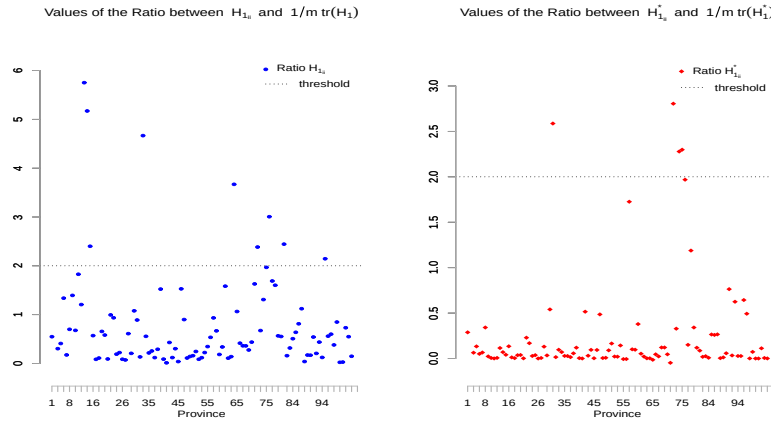


Fig. 2 Comparison between the diagonal values of matrices $\mathbf{H}_1/\text{tr}(\mathbf{H}_1)$ and $\mathbf{H}_1^*/\text{tr}(\mathbf{H}_1^*)$.

In Figures 1 and 2 we observe some differences in the patterns of dispersion points from $\mathbf{H}_1$ and $\mathbf{H}_1^*$, as there are less outlying observations on the right-hand side plot. Regarding the detection of the observations having a remarkable effect on the estimation of the model variance, matrix $\mathbf{H}_1$ captures the influential observations through the projection matrix of $\mathbf{X}$ with respect to $\mathbf{V}^{-1}$ which is, in turn, affected by the same outlying observations. Instead $\mathbf{H}_1^*$ suffers less this effect and is more useful to detect the observations that are influential for the estimation of the whole set of model parameters. We may interpret this effect by saying that outlying observations

in the explanatory variables space $\mathbf{X}$ will tend to have higher leverage, so that $\mathbf{H}_1$ could be more affected by the correlation of $\mathbf{X}$ and $\mathbf{y}$. This can be noted on the estimation of $\mathbf{V}^{-1}$, of which $\mathbf{H}_1$ is function. Since $\mathbf{H}_1^*$ is a function of the derivative of $\mathbf{H}_1$ with respect to the direct estimates $\mathbf{y}$, it gives more information about the influence of the observations on the fixed effects when the model variance is estimated. Figure 2 detects potentially influential observations (above the threshold dashed). In the left-hand plot we get twice as much influential points that in the right-hand plot. We interpret this fact by noting that some observations might be influential for the estimation of the regression parameters when the model variance is known and they are no more influential when, in addition, the variance is estimated. So some compensation process is somehow acting when fitting the data to the model.

7 Concluding remarks

We introduce the leverage matrix for the Fay-Herriot area-level model when we deal with the restricted maximum likelihood estimate of the model variance. The problem of the leverage of observed values on predicted values by the EBLUP arise when we consider that, although we utilize convenient estimates, the latter depend on influential values itself. So the leverage matrix of the model can be affected by influential observations through the estimates of the model variance. Detailed formulae on fixed and random effects leverage matrices are presented. These relations highlight the contribution of the standard leverage matrices in case of fixed $\mathbf{V}$. Influence analysis on mean squared error estimates are further investigated. They are based on $m \times m^2$ -order matrices, that can be very useful in assessing contribution of single observations (the small area direct estimates) in the evaluation of the mean squared error of all areas.

References

- Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics* **19**, 15-18.
- Demidenko, E. (2005). Influence analysis for linear mixed-effects models, *Statistics in Medicine*, **24**, 893-909.
- Fung, W.K., Zhu, Z.Y., Wei, B.C., and He, X. (2002). Influence diagnostics and outliers tests for semiparametric mixed models, *Journal of the Royal Statistical Society, Series B*, **64**, 565-579.
- Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.
- Wei, B.C., Hu, Y.Q., and Fung, W.K. (1998). Generalized leverage and its applications, *Scandinavian Journal of Statistics*, **25**, 25-37.