SUPPLEMENT TO "THE GRADIENT TEST STATISTIC FOR OUTLIER DETECTION IN GENERALIZED ESTIMATING EQUATIONS"

FELIPE OSORIO, ÁNGELO GÁRATE, AND CIBELE RUSSO

Abstract. In this supplementary material, we present details about the gradient statistic for inference functions, a brief description of the estimation procedure for generalized estimating equations (GEE), proofs of the theoretical results, tables and figures regarding to the example of the main paper and an additional real data analysis.

Appendix A. About the gradient test statistics for inference functions

A.1. Asymptotic equivalence between gradient-type and score-type test statistics.

Proposition A.1. *The gradient-type statistic given in Equation (2) from the manuscript and the score-type test statistic defined by Rotnitzky and Jewell (1990),*

$$R_n = \frac{1}{n}\boldsymbol{\Psi}_n^\top(\boldsymbol{\theta}_0)\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0),$$

are asymptotically equivalent under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Their common asymptotic distribution is $\chi^2(p)$.

Proof. To establish the asymptotic equivalence between the gradient-type and score-type statistics, we can use some results available in Yuan and Jennrich (1998). In fact,

$$\frac{1}{\sqrt{n}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) \xrightarrow{D} \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\theta}_0)),$$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{S}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{V}(\boldsymbol{\theta}_0)\boldsymbol{S}^{-\top}(\boldsymbol{\theta}_0)).$$

Now, consider the Taylor expansion of $\boldsymbol{\Psi}_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, i.e.

$$\boldsymbol{\Psi}_n(\boldsymbol{\theta}) \stackrel{\mathsf{a}}{=} \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) + \frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big|_{\theta=\theta_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where $\stackrel{\mathsf{a}}{=}$ denotes asymptotic equivalence. Evaluating at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$ and using that $\boldsymbol{\Psi}_n(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$, yields

$$\Big(-\frac{1}{n}\frac{\partial \boldsymbol{\Psi}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big|_{\theta=\theta_0}\Big)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \stackrel{\mathsf{a}}{=} \frac{1}{\sqrt{n}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0),$$

because of the additivity of the inference function in Equation (1) from the manuscript, we obtain $-\frac{1}{n}\partial\boldsymbol{\Psi}_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\top|_{\theta=\theta_0} \xrightarrow{\text{a.s.}} \boldsymbol{S}(\boldsymbol{\theta}_0)$. Henceforth, it follows

$$\boldsymbol{S}(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \stackrel{\text{a}}{=} \frac{1}{\sqrt{n}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0).$$

Thus, under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, it is straightforward to note that

$$\begin{aligned} T_n &:= \frac{1}{\sqrt{n}}\boldsymbol{\Psi}_n^\top(\boldsymbol{\theta}_0)\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\sqrt{n}\boldsymbol{S}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ &\stackrel{\text{a}}{=} \frac{1}{\sqrt{n}}\boldsymbol{\Psi}_n^\top(\boldsymbol{\theta}_0)\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\frac{1}{\sqrt{n}}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) \\ &\stackrel{\text{a}}{=} R_n, \end{aligned}$$

where $R_n = \frac{1}{n}\boldsymbol{\Psi}_n^\top(\boldsymbol{\theta}_0)\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0)$ corresponds to the score-type statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ (see Boos, 1992). This allows us to note that T_n is asymptotically equivalent to R_n and therefore, under H_0, we have $T_n \xrightarrow{D} \chi^2(p)$. □

A.2. **Gradient-type statistic for hypotheses about subvectors.** Consider that the parameter vector is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$ and suppose that we have interest in testing the hypothesis

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0, \qquad \text{against} \qquad H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^0,$$

where $\boldsymbol{\theta}_1 \in \mathbb{R}^r$ and $\boldsymbol{\theta}_2 \in \mathbb{R}^{p-r}$. Thus, assume the following partition of matrices $\boldsymbol{S}(\boldsymbol{\theta})$ and $\boldsymbol{V}(\boldsymbol{\theta})$,

$$\boldsymbol{S}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{S}_{11}(\boldsymbol{\theta}) & \boldsymbol{S}_{12}(\boldsymbol{\theta}) \\ \boldsymbol{S}_{21}(\boldsymbol{\theta}) & \boldsymbol{S}_{22}(\boldsymbol{\theta}) \end{pmatrix}, \qquad \boldsymbol{V}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{V}_{11}(\boldsymbol{\theta}) & \boldsymbol{V}_{12}(\boldsymbol{\theta}) \\ \boldsymbol{V}_{21}(\boldsymbol{\theta}) & \boldsymbol{V}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

The following lemma, adapted from Lemonte (2016), presents the gradient-type test statistic for hypotheses about subvectors.

Lemma A.2. *Under assumptions of Theorem 5.2 in Lemonte (2016) and under $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$, we have that*

$$T_n = \boldsymbol{\Psi}_1^\top(\widetilde{\boldsymbol{\theta}})\boldsymbol{Q}_{11}(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0) \xrightarrow{D} \chi^2(r), \tag{A.1}$$

where $\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = (\boldsymbol{\Psi}_1^\top(\boldsymbol{\theta}), \boldsymbol{\Psi}_2^\top(\boldsymbol{\theta}))^\top$ with $\boldsymbol{\Psi}_1(\boldsymbol{\theta})$ an r-dimensional inference function associated to $\boldsymbol{\theta}_1$, $\boldsymbol{Q}_{11}(\boldsymbol{\theta}) = \boldsymbol{V}_{11\cdot2}^{-1}(\boldsymbol{\theta})\big(\boldsymbol{S}_{11}(\boldsymbol{\theta}) - \boldsymbol{V}_{12}(\boldsymbol{\theta})\boldsymbol{V}_{22}^{-1}(\boldsymbol{\theta})\boldsymbol{S}_{21}(\boldsymbol{\theta})\big)$ and $\boldsymbol{V}_{11\cdot2}(\boldsymbol{\theta}) = \boldsymbol{V}_{11}(\boldsymbol{\theta}) - \boldsymbol{V}_{12}(\boldsymbol{\theta})\boldsymbol{V}_{22}^{-1}(\boldsymbol{\theta})\boldsymbol{V}_{21}(\boldsymbol{\theta})$.

Appendix B. Estimation in GEE

The system of equations defined in Equation (3) from the main paper typically is solved through iteratively weighted least squares (IWLS) with working response $\boldsymbol{Z}_i = \boldsymbol{F}_i\boldsymbol{\beta} + \boldsymbol{Y}_i - \boldsymbol{\mu}_i$. Thus, the estimate of $\boldsymbol{\beta}$ is updated as

$$\boldsymbol{\beta}^{(r+1)} = \Big(\sum_{i=1}^n \boldsymbol{F}_i^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i\Big)^{-1}\sum_{i=1}^n \boldsymbol{F}_i^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i,$$

where the working response, the local model matrix and the working covariance matrix $\boldsymbol{\Sigma}_i$ must be evaluated at the current estimation $\boldsymbol{\beta}^{(r)}$. Let $\boldsymbol{W}_i = \boldsymbol{D}_i^{-1}\boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}(\boldsymbol{\alpha})$

$\boldsymbol{A}_i^{-1/2}\boldsymbol{D}_i^{-1}$ and $\boldsymbol{Z}_i^* = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{D}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, for $i = 1, \ldots, n$. Thus, we can write

$$\begin{aligned}\boldsymbol{\beta}^{(r+1)} &= \Big(\sum_{i=1}^n \boldsymbol{X}_i^\top \boldsymbol{W}_i \boldsymbol{X}_i\Big)^{-1} \sum_{i=1}^n \boldsymbol{X}_i^\top \boldsymbol{W}_i \boldsymbol{Z}_i^* \\ &= (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{Z}^*,\end{aligned}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$, $\boldsymbol{W} = \oplus_{i=1}^n \boldsymbol{W}_i$ and $\boldsymbol{Z}^* = (\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top)^\top$. Note that this procedure should be alternated with the estimation of ϕ and $\boldsymbol{\alpha}$ using the method of moments (see Liang and Zeger, 1986). Considering not very restrictive conditions of regularity, Liang and Zeger (1986) showed that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ has an asymptotic normal distribution with zero mean and covariance matrix defined as

$$\boldsymbol{G}^{-1}(\boldsymbol{\beta}) = \lim_{n\to\infty} n\boldsymbol{S}_n^{-1}(\boldsymbol{\beta})\boldsymbol{V}_n(\boldsymbol{\beta})\boldsymbol{S}_n^{-1}(\boldsymbol{\beta}),$$

where

$$\boldsymbol{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{F}_i^\top \boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i, \qquad \boldsymbol{V}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{F}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathsf{Cov}(\boldsymbol{Y}_i)\boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i.$$

Liang and Zeger (1986) also showed that the empirical estimator of covariance matrix for $\widehat{\boldsymbol{\beta}}$, let say,

$$\mathsf{Cov}_{\mathsf{Emp}}(\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{G}}^{-1} = \widehat{\boldsymbol{S}}_n^{-1}\widehat{\boldsymbol{V}}_n\widehat{\boldsymbol{S}}_n^{-1},$$

with,

$$\widehat{\boldsymbol{S}}_n = \widehat{\phi}\sum_{i=1}^n \boldsymbol{X}_i^\top \widehat{\boldsymbol{W}}_i\boldsymbol{X}_i, \qquad \widehat{\boldsymbol{V}}_n = \widehat{\phi}^2\sum_{i=1}^n \boldsymbol{X}_i^\top \widehat{\boldsymbol{W}}_i\widehat{\boldsymbol{r}}_i\widehat{\boldsymbol{r}}_i^\top \widehat{\boldsymbol{W}}_i\boldsymbol{X}_i,$$

where $\boldsymbol{r}_i = \boldsymbol{D}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, for $i = 1, \ldots, n$, is a consistent estimator of $\boldsymbol{G}^{-1}(\boldsymbol{\beta})$. When the working correlation matrix is correctly specified, the model based estimator of the covariance matrix for $\widehat{\boldsymbol{\beta}}$ is given by:

$$\mathsf{Cov}_{\mathsf{MB}}(\widehat{\boldsymbol{\beta}}) = \widehat{\phi}^{-1}\Big(\sum_{i=1}^n \boldsymbol{X}_i^\top \widehat{\boldsymbol{W}}_i\boldsymbol{X}_i\Big)^{-1} = \widehat{\phi}^{-1}(\boldsymbol{X}^\top \widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}.$$

Appendix C. Main results

C.1. **One-step approximation for the estimates under the mean-shift outlier model.**

Proposition C.1. *Let $\widehat{\gamma}_i$ and $\widehat{\boldsymbol{\beta}}_*$ be the estimates of $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}$ for the mean-shift outlier model given in Equation (4) from manuscript. Based on the initial estimates $\widehat{\boldsymbol{\gamma}}_i^0 = \boldsymbol{0}$ and $\widehat{\boldsymbol{\beta}}_*^0 = \widehat{\boldsymbol{\beta}}$, we have the following one-step approximations:*

$$\widehat{\boldsymbol{\gamma}}_i^1 = (\boldsymbol{B}_i^\top \boldsymbol{W}_i(\boldsymbol{I} - \boldsymbol{H}_i)\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^\top \boldsymbol{W}_i\boldsymbol{r}_i, \tag{C.1}$$

$$\widehat{\boldsymbol{\beta}}_*^1 = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top \boldsymbol{W}_i\boldsymbol{B}_i(\boldsymbol{B}_i^\top \boldsymbol{W}_i(\boldsymbol{I} - \boldsymbol{H}_i)\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^\top \boldsymbol{W}_i\boldsymbol{r}_i, \tag{C.2}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$ and $\boldsymbol{W} = \bigoplus_{i=1}^n \boldsymbol{W}_i$ being a block diagonal matrix, with $\boldsymbol{W}_i = \boldsymbol{D}_i^{-1}\boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}(\boldsymbol{\alpha})\boldsymbol{A}_i^{-1/2}\boldsymbol{D}_i^{-1}$, $\boldsymbol{H}_i = \boldsymbol{X}_i(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top \boldsymbol{W}_i$, and $\boldsymbol{r}_i = \boldsymbol{D}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, for $i = 1, \ldots, n$.

Proof. We use one single step of the Newton-scoring algorithm (Jørgensen and Knudsen, 2004) considering the following initial estimates $\widehat{\boldsymbol{\delta}}_i^0 = (\widehat{\boldsymbol{\gamma}}_i^{0\top}, \widehat{\boldsymbol{\beta}}_*^{0\top})^\top = (\boldsymbol{0}, \widehat{\boldsymbol{\beta}}^\top)^\top$. Thus, we can write

$$\begin{aligned}\begin{pmatrix}\widehat{\boldsymbol{\gamma}}_i^1\\ \widehat{\boldsymbol{\beta}}_*^1\end{pmatrix} &= \begin{pmatrix}\boldsymbol{0}\\ \widehat{\boldsymbol{\beta}}\end{pmatrix} + \begin{pmatrix}\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i & \boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{X}_i\\ \boldsymbol{X}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i & \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{D}_i(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)\\ \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{D}(\boldsymbol{Y}-\boldsymbol{\mu})\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i & \boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{X}_i\\ \boldsymbol{X}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i & \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{Z}_i^*\\ \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{Z}^*\end{pmatrix},\end{aligned} \tag{C.3}$$

where $\boldsymbol{Z}^* = \boldsymbol{\eta} + \boldsymbol{D}(\boldsymbol{Y}-\boldsymbol{\mu})$, $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{Z}^*$. Using results of inverses for partitioned matrices and noting that

$$\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i - \boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{X}_i(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i = \boldsymbol{B}_i^\top\boldsymbol{W}_i(\boldsymbol{I}-\boldsymbol{H}_i)\boldsymbol{B}_i,$$

with $\boldsymbol{H}_i = \boldsymbol{X}_i(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top\boldsymbol{W}_i$. The aforementioned allows us to arrive at

$$\widehat{\boldsymbol{\gamma}}_i^1 = (\boldsymbol{B}_i^\top\boldsymbol{W}_i(\boldsymbol{I}-\boldsymbol{H}_i)\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^\top\boldsymbol{W}_i(\boldsymbol{Z}_i^* - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}).$$

Noting that $\boldsymbol{r}_i = \boldsymbol{Z}_i^* - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} = \boldsymbol{D}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, we obtain (C.1). Besides, from the second equation in (C.3), it follows that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_*^1 &= \widehat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i(\boldsymbol{B}_i^\top\boldsymbol{W}_i(\boldsymbol{I}-\boldsymbol{H}_i)\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}\\ &\quad - (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i(\boldsymbol{B}_i^\top\boldsymbol{W}_i(\boldsymbol{I}-\boldsymbol{H}_i)\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{Z}_i\\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top\boldsymbol{W}_i\boldsymbol{B}_i(\boldsymbol{B}_i^\top\boldsymbol{W}_i(\boldsymbol{I}-\boldsymbol{H}_i)\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^\top\boldsymbol{W}_i\boldsymbol{r}_i,\end{aligned}$$

which makes it possible to verify (C.2). □

Remark C.2. An one-step approximation for the gradient-type test statistic, denoted as T_i^1, can be obtained using estimates given in Equations (C.1) and (C.2).

C.2. **Connection between gradient and generalized Cook distances for inference functions.** First note that, due to the additivity of $\boldsymbol{\Psi}_n(\boldsymbol{\theta})$ defined in Equation (1) from manuscript, we have that

$$\boldsymbol{0} = \boldsymbol{\Psi}_n(\widehat{\boldsymbol{\theta}}) = \sum_{j\neq i}^n \boldsymbol{\Psi}_j(\widehat{\boldsymbol{\theta}}) + \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\Psi}_{(i)}(\widehat{\boldsymbol{\theta}}) + \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}),$$

this leads to $\boldsymbol{\Psi}_{(i)}(\widehat{\boldsymbol{\theta}}) = -\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}})$. Thus, we obtain the following one-step approximation

$$\widehat{\boldsymbol{\theta}}_{(i)}^1 = \widehat{\boldsymbol{\theta}} - \boldsymbol{S}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}). \tag{C.4}$$

It is noteworthy that by (C.4), we can write

$$\begin{aligned}TD_i &\approx \boldsymbol{\Psi}_i^\top(\widehat{\boldsymbol{\theta}})\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{S}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}^1)\\ &= \boldsymbol{\Psi}_i^\top(\widehat{\boldsymbol{\theta}})\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}) = GD_i,\end{aligned}$$

where GD_i is known as the generalized Cook distance for the inference functions framework. That is, TD_i^1 defined as,

$$TD_i^1 = \boldsymbol{\Psi}_i^\top(\widehat{\boldsymbol{\theta}})\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{S}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}^1),$$

corresponds to a one-step approximation for TD_i, $i = 1, \dots, n$.

Remark C.3. It is easy to see that we can naturally extend the gradient distance TD_I, to evaluate the effect of a subset of m observations, $I = \{i_1, i_2, \ldots, i_m\}$, with appropriate modifications into the notation of the expressions defined in Section 3.2 from manuscript. Moreover, our proposal for the gradient distance defined in Equation (6) generalises the influence measure introduced by Enea and Plaia (2017). In fact, under the maximum likelihood framework, we have that $\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{F}}(\boldsymbol{\theta})$, with $\boldsymbol{\mathcal{F}}(\boldsymbol{\theta})$ the Fisher information matrix, which leads to

$$TD_i = \boldsymbol{\Psi}_i^\top(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}), \qquad i = 1, \ldots, n,$$

whereas in that case $\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}})$ corresponds to the score function associated to the ith subject.

C.3. **Equivalence of case-deletion and mean-shift outlier model in GEE.** In the following we focus on estimation in the mean-shift outlier model as well as on the case-deletion method in the setting of GEE. Let $\widehat{\boldsymbol{\beta}}_{(i)}$ be the estimate of $\boldsymbol{\beta}$ when the ith subject (i.e. cluster) is removed from the dataset. The following proposition establish the equivalence of the coefficient estimates for the case-deletion and mean-shift outlier model in GEE.

Proposition C.4. *For the mean-shift outlier model established in Equation (4) of the manuscript with $\boldsymbol{B}_i = \boldsymbol{I}_{n_i}$, we have that $\widehat{\boldsymbol{\beta}}^1_* = \widehat{\boldsymbol{\beta}}^1_{(i)}$, where $\widehat{\boldsymbol{\beta}}^1_*$ is given in (C.2) and $\widehat{\boldsymbol{\beta}}^1_{(i)}$ denotes the one-step approximation of $\widehat{\boldsymbol{\beta}}_{(i)}$.*

Proof. Consider the elements defined in Proposition C.1. When $\boldsymbol{B}_i = \boldsymbol{I}_{n_i}$, we have that the one-step approximation given in (C.2), adopts the form:

$$\widehat{\boldsymbol{\beta}}^1_* = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}_i^\top \boldsymbol{W}_i (\boldsymbol{I} - \boldsymbol{H}_i)^{-1} \boldsymbol{r}_i.$$

By Corollary 1.1 of Preisser and Qaqish (1996) and noting that $\boldsymbol{W}_i(\boldsymbol{I} - \boldsymbol{H}_i)^{-1} = (\boldsymbol{W}_i^{-1} - \boldsymbol{Q}_i)^{-1}$ with $\boldsymbol{Q}_i = \boldsymbol{X}_i(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}_i^\top$, follows that $\widehat{\boldsymbol{\beta}}^1_{(i)} = \widehat{\boldsymbol{\beta}}^1_*$ as desired. □

Appendix D. Additional results

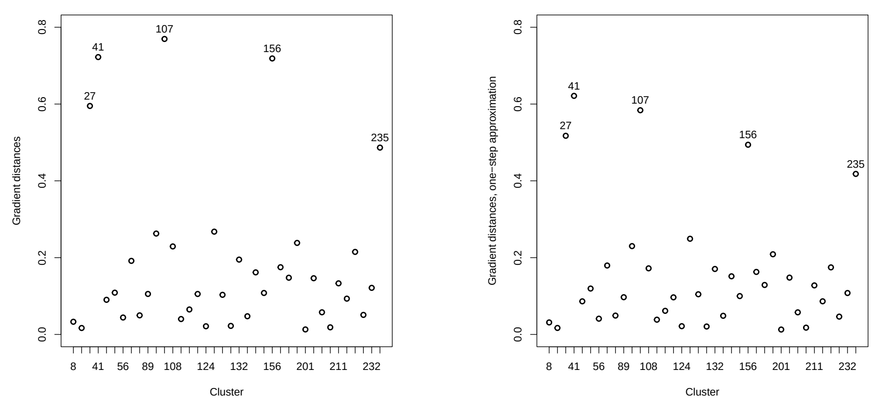
Table D.1. Parameter estimates, standard errors (in parenthesis), p-values and percentage change in parameter estimates for GUIDE data using GEE.

Variable	Full data	Removed observations					
		8	44	64	86	88	122
Intercept	-3.054	-2.656	-3.365	-3.405	-3.088	-3.504	-3.180
	(0.959)	(0.823)	(1.028)	(1.018)	(0.958)	(0.992)	(1.018)
	0.001	0.001	0.001	0.001	0.001	0.000	0.002
	—	-13.0%	10.2%	11.5%	1.1%	14.7%	4.1%
sex	-0.745	-1.077	-0.760	-0.501	-0.735	-0.767	-0.748
	(0.600)	(0.560)	(0.637)	(0.575)	(0.602)	(0.625)	(0.581)
	0.214	0.054	0.233	0.384	0.222	0.220	0.198
	—	44.5%	2.0%	-32.8%	-1.3%	2.9%	0.3%

(continued)

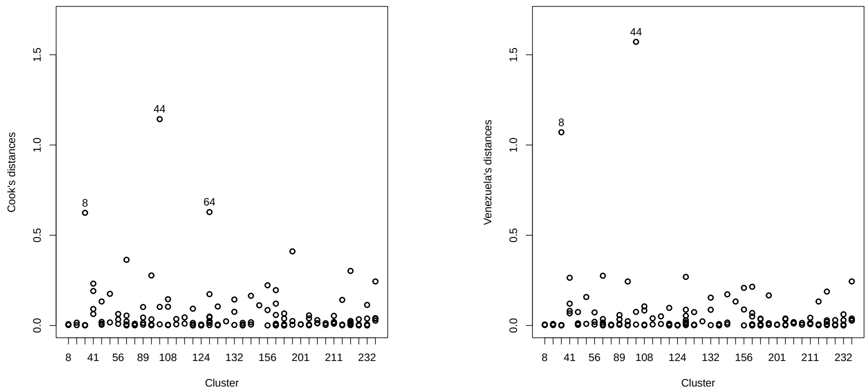
Variable	Full data	Removed observations					
		8	44	64	86	88	122
age	-0.676	-0.913	-0.782	-0.546	-0.671	-0.742	-0.698
	(0.561)	(0.577)	(0.594)	(0.588)	(0.562)	(0.567)	(0.550)
	0.228	0.114	0.188	0.353	0.233	0.190	0.204
	—	35.1%	15.8%	-19.2%	-0.7%	9.9%	3.4%
accidents	0.392	0.458	0.391	0.408	0.395	0.427	0.376
	(0.093)	(0.099)	(0.101)	(0.098)	(0.099)	(0.098)	(0.091)
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	—	16.9%	-0.2%	4.2%	0.8%	8.9%	-4.2%
severe	0.812	0.646	0.722	0.823	0.815	0.956	0.944
	(0.359)	(0.330)	(0.351)	(0.357)	(0.360)	(0.367)	(0.395)
	0.024	0.050	0.039	0.021	0.024	0.009	0.017
	—	-20.5%	-11.1%	1.4%	0.3%	17.7%	16.2%
toilet	0.108	0.143	0.209	0.104	0.108	0.112	0.096
	(0.099)	(0.117)	(0.100)	(0.096)	(0.099)	(0.105)	(0.097)
	0.276	0.219	0.037	0.278	0.275	0.290	0.322
	—	32.9%	93.9%	-3.1%	0.5%	3.9%	-10.5%
$\widehat{\alpha}$	0.093	0.086	0.102	0.103	0.078	0.031	0.104
	—	-7.4%	9.5%	10.7%	-15.7%	-66.3%	11.1%
$\det(\mathsf{Cov}_{\mathsf{Emp}}(\widehat{\boldsymbol{\beta}}))^1$	0.300	0.349	0.312	0.448	0.352	0.392	0.321
	—	16.1%	4.0%	49.0%	17.1%	30.3%	7.0%
$\det(\mathsf{Cov}_{\mathsf{MB}}(\widehat{\boldsymbol{\beta}}))^1$	0.488	0.862	0.825	0.572	0.532	0.732	0.557
	—	76.5%	68.9%	17.2%	9.0%	49.8%	14.1%

[1] values multiplied by 10^7.



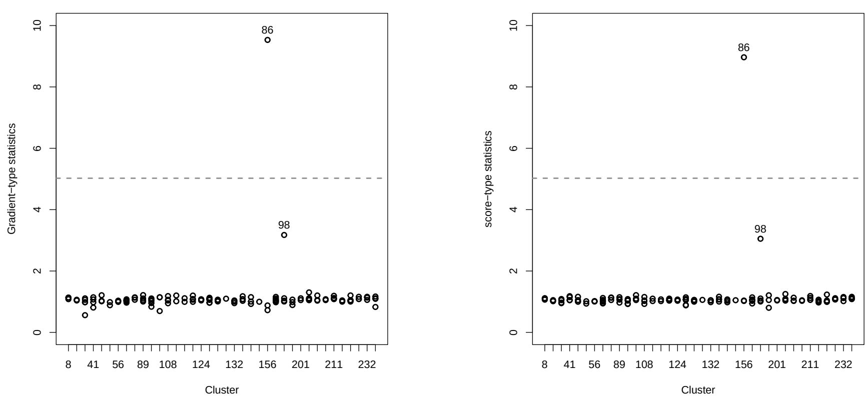
(a) (b)

FIGURE D.1. Cluster-level gradient distance for GUIDE data: (a) full iterated and (b) one-step aproximation.



(a) (b)

Figure D.2. Observation-level influence measures for GUIDE data: one-step approximation of (a) Cook's distances D_{ij}^{1} and (b) Venezuela's distances $D_{ij,\mathsf{VBS}}^{1}$.



(a) (b)

Figure D.3. Observation-level one-step approximations of test statistics the null hipothesis $H_0 : \gamma_i = 0$ for GUIDE data: (a) gradient-type T_{ij}^{1} and (b) score-type R_{ij}^{1} statistics.

Appendix E. Real data example

Lima and Sañudo (1997) conducted a study to evaluate the learning process in a certain manual task in which 40 volunteer subjects participated, who had 8 attempts to perform the same task. The time difference between a subject received a stimulus and the instant that a motor reaction occurred (in milliseconds) was recorded. The observed data for each of the subjects (clusters) are displayed in Figure E.1

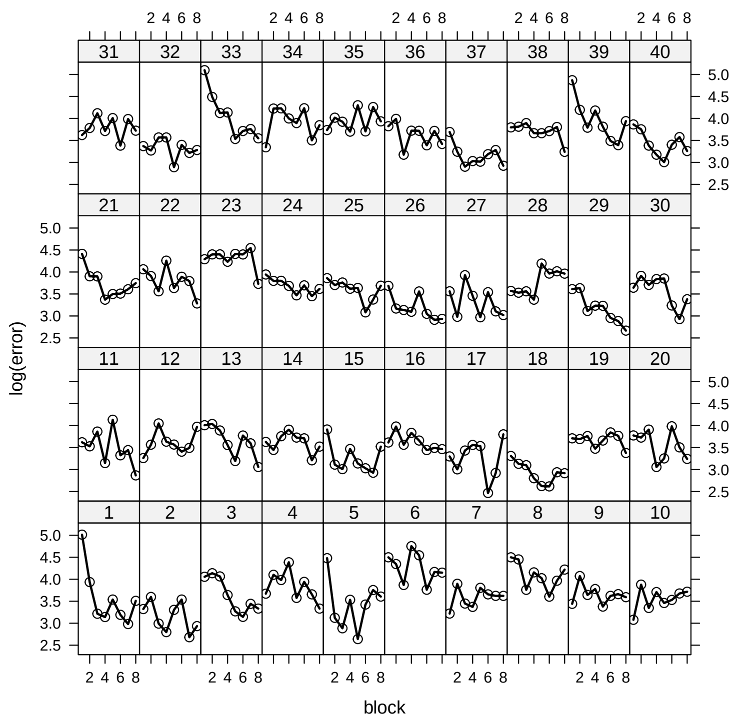


Figure E.1. Logarithm of the observed time difference between attempts to perform a task, Lima and Sañudo (1997) dataset.

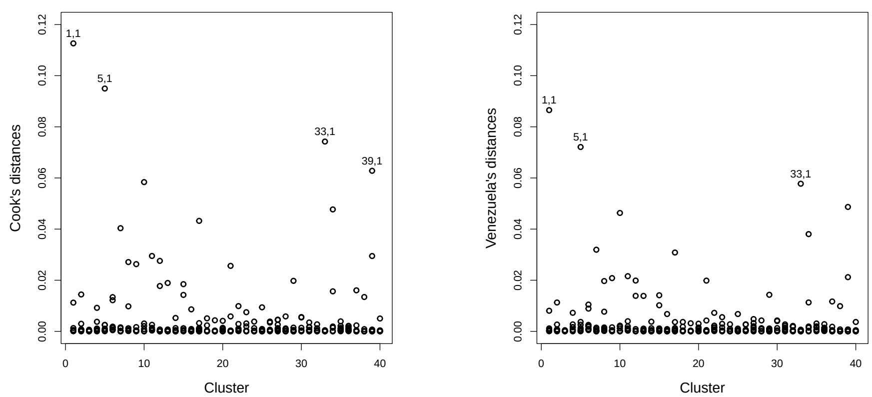
This dataset has been previously analyzed by Venezuela et al. (2007, 2011) who assumed a Gaussian model with identity link function for the logarithm of the response variable considering,

$$\mu_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}, \qquad \boldsymbol{x}_{ij} = (1, j)^{\top}, \qquad \boldsymbol{\beta} = (\beta_0, \beta_1)^{\top},$$

for $i = 1, \dots, 40; j = 1, \dots, 8$. The model was fitted using GEE with working correlation matrix following AR(1) structure. Figure E.2 displays the Cook's distances based on Equations (7) and (8) from the main manuscript. Although this graph suggests that the first attempt of individuals 1, 5, 33 and 39 are influential observations, the confirmatory analysis reported in Table E.1 reveals that their effect is quite marginal and they do not produce any inferential change.

TABLE E.1. Parameter estimates, standard errors (in parenthesis), p-values and percentage change in parameter estimates for Lima and Sañudo data using GEE.

Parameter	Full data	Removed observations				
		1,1	5,1	5,5	33,1	39,1
β_0	3.8500	3.8299	3.8317	3.8519	3.8335	3.8349
	(0.0665)	(0.0644)	(0.0694)	(0.0667)	(0.0613)	(0.0640)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	—	-0.52%	-0.48%	0.05%	-0.43%	-0.39%
β_1	-0.0513	-0.0481	-0.0483	-0.0511	-0.0487	-0.0489
	(0.0101)	(0.0095)	(0.0104)	(0.0101)	(0.0094)	(0.0099)
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	—	-6.27%	-5.77%	-0.26%	-5.08%	-4.67%
ϕ^{-1}	0.1730	0.1687	0.1720	0.1707	0.1681	0.1698
	—	-2.47%	-0.58%	-1.33%	-2.84%	-1.84%
α	0.5309	0.5400	0.5440	0.5342	0.5269	0.5310
	—	1.71%	2.46%	0.61%	-0.76%	0.00%



(a) (b)

FIGURE E.2. Observation-level influence measures for Lima and Sañudo data: (a) Cook's distances and (b) Venezuela's distances.

We should highlight the ability of the gradient-type statistic to determine that there are no outliers in this dataset (see Figure E.3). Indeed, this is in agreement with the results presented in Venezuela et al. (2007), who through a simulated envelope reach the same conclusion. Additionally, we stress that the performance of the gradient-type statistic is very similar to the type-score statistic (results not shown here).

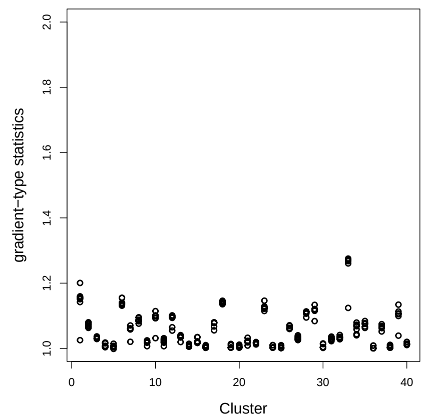


FIGURE E.3. Observation-level gradient-type test statistics of null hipothesis $H_0 : \gamma_i = 0$ for Lima and Sañudo data.

REFERENCES

Boos, D.D. (1992). On generalized score tests. *The American Statistician* **46**, 327-333.

Crudu, F., Osorio, F. (2020). Bilinear form test statistics for extremum estimation. *Economics Letters* **187**, 108885.

Enea, M., Plaia, A. (2017). A gradient-based deletion diagnostic measure for generalized linear mixed models. *Communications in Statistics - Theory and Methods* **46**, 1972-1982.

Jørgensen, B., Knudsen, S.J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics* **31**, 93-114.

Lemonte, A.J. (2016). *The Gradient Test: Another Likelihood-based Test*. Academic Press, London.

Liang, K.-Y., Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lima, A.C.P., Sañudo, A. (1997). Transferência entre tarefas sincronizatórias com diferentes níveis de complexidade. Technical Report RAE-CEA-9702. IME-USP, São Paulo.

Preisser, J.S., Qaqish, B.F. (1996). Deletion diagnostics for generalized estimating equations. *Biometrika* **83**, 551-562.

Rotnitzky, A., Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered data. *Biometrika* **77**, 485-497.

Venezuela, M.K., Botter, D.A., Sandoval, M.C. (2007). Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation* **77**, 879-888.

Venezuela, M.K., Sandoval, M.C., Botter, D.A. (2011). Local influence in estimating equations. *Computational Statistics and Data Analysis* **55**, 1867-1883.

Yuan, K.H., Jennrich, R.I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis* **65**, 245-260.

Departamento de Matemática, Universidad Técnica Federico Santa María, Chile
Current address: Avenida España 1680, Valparaíso, Chile
Orcid ID: http://orcid.org/0000-0002-4675-5201
Email address: felipe.osorios@usm.cl

Departamento de Estadística, Pontificia Universidad Católica de Chile
Current address: Avenida Vicuña Mackena 4860, Santiago, Chile
Email address: afgarate@uc.cl

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil
Current address: Av. Trabalhador São Carlense 400, São Carlos, Brazil.
Orcid ID: http://orcid.org/0000-0003-1356-0245
Email address: cibele@icmc.usp.br