Robust semiparametric inference for polytomous logistic regression with complex survey design

Castilla, E.[1]; Ghosh, A[2]; Martin, N.[1] and Pardo, L.[1]
[1]Complutense University of Madrid, Madrid, Spain
[2]Indian Statistical Institute, Kolkata, India

Abstract

Analyzing polytomous response from a complex survey scheme, like stratified or cluster sampling is very crucial in several socio-economics applications. We present a class of minimum quasi weighted density power divergence estimators for the polytomous logistic regression model with such a complex survey. This family of semiparametric estimators is a robust generalization of the maximum quasi weighted likelihood estimator exploiting the advantages of the popular density power divergence measure. Accordingly robust estimators for the design effects are also derived. Robust testing of general linear hypotheses on the regression coefficients are proposed using the new estimators. Their asymptotic distributions and robustness properties are theoretically studied and also empirically validated through a numerical example and an extensive Monte Carlo study.

Keywords: Cluster sampling; Design effect; Minimum quasi weighted DPD estimator; Polytomous logistic regression model; Pseudo minimum phi-divergence estimator; Quasi-likelihood; Robustness

1 Introduction

In many real-life applications, we come across data that have been collected through a complex survey scheme, like stratified sampling or cluster sampling, etc., rather than the simple random sampling. Such situations commonly arise in large scale data collection, for example, within several states of a country or even among different countries. Suitable statistical methods are required to analyze these data by taking care of the stratified structure of the data; this is because there often exist several inter and intra-class correlations within such stratification and ignoring them often lead to erroneous inference. Further, in many such complex surveys, stratified observations are collected on some categorical responses having two or more mutually exclusive unordered categories along with some related covariates and inference about their relationship is of up-most interest for insight generation and policy making. Polytomous logistic regression (PLR) model is a useful and popular tool in such situations to model categorical responses with associated covariates. However, most of classical literature deals with the cases of simple random sampling scheme (e.g. McCullagh, 1980; Lesaffre and Albert, 1989; Agresti, 2002; Gupta et al., 2006, 2008). The application of PLR model

under complex survey setting can be found, for example, in Binder (1983), Roberts et al. (1987), Morel (1989), Morel and Neerchal (2012) and Castilla et al. (2018); most of them, except the last one, are based on the quasi maximum likelihood approach.

Even though the maximum quasi weighted likelihood estimator is the base of most of the existing literature on logistic models under complex survey designs, it is known to be non-robust in the presence of possible outliers in the data. In practice, with such a complex survey design, it is quite natural to have some outlying observations that make the likelihood based inference highly unstable. So, we often may need to make additional efforts to find and discard the outliers from the data before their analyses. A robust method providing stable solution even in presence of the outliers will be really helpful and more efficient in practice. The cited work by Castilla et al (2018) has developed an alternative minimum divergence estimator based on ϕ-divergences (Pardo, 2006), but the important issue of robustness is still ignored there.

In this paper, we develop a robust estimator under the PLR model with a complex survey based on a minimum quasi weighted divergence approach. In particular, we exploit the nice properties of the density power divergence (DPD) of Basu et al. (1998). This measure become very popular in recent literature for yielding highly robust and efficient estimators under various statistical models; see, for example, Ghosh and Basu (2013, 2016, 2018) and, in particular, the recent paper by Castilla et al (2019) which discussed a PLR model but under the simple random sampling scheme.

We first start with the mathematical description of the PLR model with complex survey set-up and a brief discussion about the maximum quasi weighted likelihood estimator of the underlying parameters in Section 2. Then we introduce a class of new robust parameter estimates for the PLR model with complex survey by minimizing a suitably defined DPD measure in Section 3. The asymptotic distribution of the resulting estimator and the design effect for the PLR model with complex survey are also described there. In Section 4, a new family of Wald-type tests is introduced based on our new estimators for testing linear hypotheses about the parameters of the PLR model. In Section 5 we theoretically study the robustness of the proposed estimators and Wald-type tests through the influence function analysis. After presenting an illustrative example in Section 6, an extensive simulation study is presented in Section 7. The paper ends with a brief concluding remark in Section 8. For brevity in presentation, proofs of all the results are given in Appendix A.

2 Maximum Quasi Weighted Likelihood Estimator

Let us assume that the whole population is partitioned into H distinct strata and the data consist of n_h clusters in stratum h for each $h = 1, \ldots, H$. Further, for each cluster $i = 1, \ldots, n_h$ in the stratum h, we have observed the values of a categorical response variable (Y) for m_{hi} units. Assuming Y has $(d+1)$ categories, we denote these observed responses by a $(d+1)$-dimensional classification vector

$$\boldsymbol{y}_{hij} = (y_{hij1}, ..., y_{hij,d+1})^T \,,\ h = 1, ..., H,\ i = 1, ..., n_h,\ j = 1, ..., m_{hi}, \tag{1}$$

with $y_{hijr} = 1$ and $y_{hijl} = 0$ for $l \in \{1, ..., d+1\} - \{r\}$ if the j-th unit selected from the i-th cluster of the h-th stratum falls in the r-th category. We also have data on $(k+1)$ explanatory variables which are common for all the individuals in the i-th cluster of the h-th stratum (very common with

dummy or qualitative explanatory variables) to be denoted as $\boldsymbol{x}_{hi} = (x_{hi0}, x_{hi1}, ..., x_{hik})^T$; the first one $x_{hi0} = 1$ is associated with the intercept. Let us denote the sampling weight from the i-th cluster of the h-th stratum by w_{hi}. For each i, h and j, the expectation of the r-th element of the random variable $\boldsymbol{Y}_{hij} = (Y_{hij1}, ..., Y_{hij,d+1})^T$, corresponding to the realization $\boldsymbol{y}_{hij}$, is given by the PLR model

$$\pi_{hir}(\boldsymbol{\beta}) = \mathrm{E}[Y_{hijr}|\boldsymbol{x}_{hi}] = \Pr(Y_{hijr} = 1|\boldsymbol{x}_{hi}) = \begin{cases} \dfrac{\exp\{\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_r\}}{1 + \sum_{l=1}^{d}\exp\{\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_l\}}, & r = 1, ..., d, \\ \dfrac{1}{1 + \sum_{l=1}^{d}\exp\{\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_l\}}, & r = d+1, \end{cases}, \quad (2)$$

with $\boldsymbol{\beta}_r = (\beta_{r0}, \beta_{r1}, ..., \beta_{rk})^T \in \mathbb{R}^{k+1}$, $r = 1, ..., d$. Note that, the expectation of $\boldsymbol{Y}_{hij}$ does not depends on the unit number j (homogeneity), which is not a strong assumption as we generally have random sampling with the clusters in each stratum. Let $\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})$ denote the $(d+1)$-dimensional probability vector with the elements given in (2), i.e.,

$$\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}) = (\pi_{hi1}(\boldsymbol{\beta}), ..., \pi_{hi,d+1}(\boldsymbol{\beta}))^T. \quad (3)$$

Then, the associated parameter space for (2) is given by

$$\Theta = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_d^T)^T,\ \boldsymbol{\beta}_r = (\beta_{r0}, ..., \beta_{rk})^T \in \mathbb{R}^{k+1},\ r = 1, ..., d\} = \mathbb{R}^{d(k+1)}.$$

For modeling data exhibiting overdispersion, the quasi-likelihood method is an widely used method, originally defined by Wedderburn (1974). The quasi-loglikelihood function is constructed without complete distributional knowledge but through the mean and the variance of independent sampling units, the total of clusters in all the strata $n = \sum_{h=1}^{H} n_h$ in the PLR model with complex sampling as given in (2). It is semiparametric method, since it only specifies the first two multivariate moments of $\boldsymbol{Y}_{hij}$. Let $f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ be the probability mass function of $\boldsymbol{Y}_{hij}$ such that $\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})$ is modeled by the PLR model with complex sampling, i.e. since the support of $\boldsymbol{Y}_{hij}$, say $\mathbb{Y}_{hi}$, is the set of column vectors of identity matrix $\boldsymbol{I}_{d+1}$ it holds

$$f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}) = \boldsymbol{\pi}_{hi}^T(\boldsymbol{\beta})\,\boldsymbol{y}_{hij} = \sum_{l=1}^{d}\pi_{hil}(\boldsymbol{\beta})\,y_{hijl},$$

$$\log f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}) = \log\left(\boldsymbol{\pi}_{hi}^T(\boldsymbol{\beta})\,\boldsymbol{y}_{hij}\right) = \sum_{l=1}^{d}\log\pi_{hil}(\boldsymbol{\beta})\,y_{hijl} = \log\boldsymbol{\pi}_{hi}^T(\boldsymbol{\beta})\,\boldsymbol{y}_{hij},$$

where we use the notation $\log\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}) = (\log\pi_{hi1}(\boldsymbol{\beta}), ..., \log\pi_{hi,d+1}(\boldsymbol{\beta}))^T$. It is important to be aware that the correct loglikelihood should be $\ell(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\log f_{\boldsymbol{\beta}}(\underline{\boldsymbol{y}}_{hi}|\boldsymbol{x}_{hi})$, where $\underline{\boldsymbol{y}}_{hi} = (\boldsymbol{y}_{hi1}, ..., \boldsymbol{y}_{him_{hi}})^T$. Under homogeneity assumption within the clusters, since $f_{\boldsymbol{\beta}}(\underline{\boldsymbol{y}}_{hi}|\boldsymbol{x}_{hi})$ is unknown, the quasi loglikelihood, $\ell(\boldsymbol{\beta})$, considers the likelihood within each cluster the same as independent case as an approximation,

$$\log f_{\boldsymbol{\beta}}(\underline{\boldsymbol{y}}_{hi}|\boldsymbol{x}_{hi}) \overset{def}{=} \sum_{j=1}^{m_{hi}}\log f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), \quad (4)$$

and thus

$$\ell(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \log \boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})\, \boldsymbol{y}_{hij} = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \log \boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})\, \widehat{\boldsymbol{y}}_{hi},$$

where

$$\widehat{\boldsymbol{y}}_{hi} = \mathbf{1}_{m_{hi}}^{T}\underline{\boldsymbol{y}}_{hi} = \sum_{j=1}^{m_{hi}} \boldsymbol{y}_{hij}, \quad \widehat{\boldsymbol{y}}_{hi} = (\widehat{y}_{hi1}, ..., \widehat{y}_{hi,d+1})^{T} = \left(\sum_{j=1}^{m_{hi}} y_{hij1}, ..., \sum_{j=1}^{m_{hi}} y_{hij,d+1}\right)^{T}, \tag{5}$$

denotes the realization value of counts in the i-th cluster of the h-th stratum.

An important feature of the quasi loglikelihood is that the marginal distributions of $\boldsymbol{Y}_{hij}$ are completely known but the components of $\underline{\boldsymbol{Y}}_{hi}$, jointly, might be correlated. This means that distribution of their total, $\widehat{\boldsymbol{Y}}_{hi}$, might be also unknown, but the expectation is obtained as the total of the expectation of $\boldsymbol{Y}_{hij}$, $j = 1, ..., m_{hi}$. The most common assumption is to consider that $\widehat{\boldsymbol{Y}}_{hi}$ has a multinomial sampling scheme, which means that $\boldsymbol{Y}_{hij}$, $j = 1, ..., m_{hi}$ are independent random variables and

$$\boldsymbol{\Sigma}_{hi} = \boldsymbol{\Sigma}_{hi}(\boldsymbol{\beta}) = m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})), \tag{6}$$

where $\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})) = \text{diag}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})) - \boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\,\boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})$. Since (4) is not an approximation, the term "quasi" should be dropped. A weaker assumption is to consider that $\widehat{\boldsymbol{Y}}_{hi}$ has a multinomial sampling scheme with a overdispersion parameter $\nu_{hi} = 1 + \rho_{hi}^{2}(m_{hi} - 1)$, which means that $\boldsymbol{Y}_{hij}$, $j = 1, ..., m_{hi}$ are correlated random variables ($Cor[\boldsymbol{Y}_{hia}, \boldsymbol{Y}_{hib}] = \rho_{hi}^{2}$, $a \neq b$, $a, b \in \{1, ..., m_{hi}\}$) and

$$\boldsymbol{\Sigma}_{hi} = \boldsymbol{\Sigma}_{hi}(\nu_{hi}, \boldsymbol{\beta}) = \nu_{hi}m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})), \tag{7}$$

but the distribution of $\widehat{\boldsymbol{Y}}_{hi}$ is not in principle used for the estimators. Distributions such as Dirichlet Multinomial, Random Clumped and m-inflated belong this family (see Alonso et al. (2017), Morel and Neerchal (2012) and Raim et al. (2015) for details). The weakest assumption is to consider that $\widehat{\boldsymbol{Y}}_{hi}$ has an unknown distribution, with $\boldsymbol{Y}_{hij}$, $j = 1, ..., m_{hi}$ being possibly correlated but with no specific pattern. It is worth of mentioning that ν_{hi} plays here a role of nuisance parameter and it is possible to consider a model with additional nuisance parameters, which are more complex than (7) but simpler than the option of completely unknown distribution for $\widehat{\boldsymbol{Y}}_{hi}$ (see Morel and Koehler (1995) for details).

Taking into account weights for each cluster, the quasi weighted loglikelihood is defined as

$$\ell(\boldsymbol{\beta}, w) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi} \log \boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})\, \widehat{\boldsymbol{Y}}_{hi}. \tag{8}$$

Then, the maximum quasi weighted likelihood estimator of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}_{P}$, is obtained by maximizing the quasi weighted loglikelihood, $\ell(\boldsymbol{\beta}, w)$, with respect to $\boldsymbol{\beta}$. The corresponding estimating equation is then given by

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}\frac{\partial \boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\text{diag}^{-1}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\left[\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\right] = \mathbf{0}_{d(k+1)}, \tag{9}$$

with

$$\frac{\partial \boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})) \otimes \boldsymbol{x}_{hi}, \quad \boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})) = (\boldsymbol{I}_d, \mathbf{0}_d)\, \boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})).$$

The system of equations (9) can be written as $\boldsymbol{u}(\boldsymbol{\beta}) = \mathbf{0}_{d(k+1)}$, where

$$\boldsymbol{u}(\boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \boldsymbol{u}_{hi}(\boldsymbol{\beta}, \boldsymbol{x}_{hi}), \quad \boldsymbol{u}(\boldsymbol{\beta}, \boldsymbol{x}_{hi}) = w_{hi} \left[\widehat{\boldsymbol{Y}}_{hi}^{*} - m_{hi} \boldsymbol{\pi}_{hi}^{*}(\boldsymbol{\beta}) \right] \otimes \boldsymbol{x}_{hi}, \tag{10}$$

with superscript * here, the vector (matrix) obtained by deleting the last row from the initial vector (matrix) is denoted; thus $\boldsymbol{\pi}_{hi}^{*}(\boldsymbol{\beta}) = (\pi_{hi1}(\boldsymbol{\beta}), ..., \pi_{hid}(\boldsymbol{\beta}))^T$ and $\widehat{\boldsymbol{Y}}_{hi}^{*} = \left(\widehat{Y}_{hi1}^{*}, ..., \widehat{Y}_{hid}^{*}\right)^T$. The derivation from (9) to (10) is in Appendix and some additional details can be found in Morel (1989).

3 The Minimum Quasi Weighted Density Power Divergence estimators

Let $f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ be the probability mass function of $\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi}$ as defined in the previous section, $g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ an unknown and true probability mass function of $\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi}$ and $\mathbb{Y}_{hi}$ the support. The DPD based on the probability mass functions of a single observation of the sample, between $f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ and $g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$, is given by, for $\lambda > 0$,

$$\begin{aligned} d_\lambda\left(g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right) &= \int_{\mathbb{Y}_{hi}} \left(f_{\boldsymbol{\beta}}^{\lambda+1}(\boldsymbol{y}|\boldsymbol{x}_{hi}) - \frac{\lambda+1}{\lambda} f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{y}|\boldsymbol{x}_{hi}) g(\boldsymbol{y}|\boldsymbol{x}_{hi}) + \frac{1}{\lambda} g^{\lambda+1}(\boldsymbol{y}|\boldsymbol{x}_{hi}) \right) d\boldsymbol{y} \\ &= \int_{\mathbb{Y}_{hi}} f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{y}|\boldsymbol{x}_{hi}) dF(\boldsymbol{y}|\boldsymbol{x}_{hi}) - \frac{\lambda+1}{\lambda} \int_{\mathbb{Y}_{hi}} f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{y}|\boldsymbol{x}_{hi}) dG(\boldsymbol{y}|\boldsymbol{x}_{hi}) + K \\ &= d_\lambda^{*}\left(g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right) + K, \end{aligned}$$

where K is a constant not depending on $\boldsymbol{\beta}$, $F(\boldsymbol{y}|\boldsymbol{x}_{hi})$ and $G(\boldsymbol{y}|\boldsymbol{x}_{hi})$ are the distribution functions corresponding to the densities $f(\boldsymbol{y}|\boldsymbol{x}_{hi})$ and $g(\boldsymbol{y}|\boldsymbol{x}_{hi})$, respectively, and

$$d_\lambda^{*}\left(g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right) = E[f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi})] - \frac{\lambda+1}{\lambda} \int_{\mathbb{Y}_{hi}} f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{y}|\boldsymbol{x}_{hi}) dG(\boldsymbol{y}|\boldsymbol{x}_{hi})$$

is the kernel of $d_\lambda\left(g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right)$. In practice, since G is unknown, it must be estimated from the sample, which is in this case a single individual, so

$$d_\lambda^{*}\left(\widehat{g}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right) = E[f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi})] - \frac{\lambda+1}{\lambda} f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}).$$

Based on Ghosh and Basu (2013), the "kernel of the ordinary DPD" between the probability mass functions $\widehat{g}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ and $f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ for the whole sample is defined as a total discrepancy given by

$$d_\lambda^{*}(\widehat{g}, f_{\boldsymbol{\beta}}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \left(E\left[f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi}) \right] - \frac{\lambda+1}{\lambda} f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}) \right), \quad \text{for } \lambda > 0.$$

Since the support of $\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi}$, is the set of column vectors of identity matrix $\boldsymbol{I}_{d+1}$, it holds

$$E\left[f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi})\right] = E\left[\boldsymbol{\pi}_{hi}^{\lambda,T}(\boldsymbol{\beta})\,\boldsymbol{Y}_{hij}\right] = \sum_{l=1}^{d+1}\pi_{hil}^{\lambda+1}(\boldsymbol{\beta}).$$

In the current paper it is defined for the first time the "kernel of the quasi weighted DPD" between the probability mass functions $\widehat{g}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ and $f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ in the whole sample as a weighted sum

$$d_{\lambda}^{*}(\widehat{g}, f_{\boldsymbol{\beta}}, w) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hi}\left(E\left[f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{Y}_{hij}|\boldsymbol{x}_{hi})\right] - \frac{\lambda+1}{\lambda}f_{\boldsymbol{\beta}}^{\lambda}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right), \quad \text{for } \lambda > 0,$$

whose expression for the PLR model with complex sampling is

$$d_{\lambda}^{*}(\widehat{g}, f_{\boldsymbol{\beta}}, w) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}\boldsymbol{\pi}_{hi}^{\lambda,T}(\boldsymbol{\beta})\left(m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}) - \frac{\lambda+1}{\lambda}\widehat{\boldsymbol{y}}_{hi}\right), \quad \text{for } \lambda > 0. \tag{11}$$

Based on $d_{\lambda}^{*}(\widehat{g}, f_{\boldsymbol{\beta}}, w)$ the minimum quasi weighted DPD estimator is formally defined as follows.

Definition 1 *The minimum quasi weighted DPD estimator, say $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$, of $\boldsymbol{\beta}$ is defined as*

$$\widehat{\boldsymbol{\beta}}_{\lambda,Q} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d(k+1)}} d_{\lambda}^{*}(\widehat{g}, f_{\boldsymbol{\beta}}, w).$$

At the particular choice $\lambda \to 0$, the DPD measure, defined as the limit of (11) coincides (in limit) with the weighted quasi-loglikelihood, $\ell(\boldsymbol{\beta}, w)$, given in (8); thus the minimum quasi weighted DPD estimator of $\boldsymbol{\beta}$ at $\lambda = 0$ coincides with the maximum weighted quasi likelihood estimator. With the same philosophy, the following result generalizes $\boldsymbol{u}(\boldsymbol{\beta}, \boldsymbol{x}_{hi})$ given in (10) which plays an important role for the derivation of the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$.

Theorem 2 *The minimum quasi weighted DPD estimate of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$, can be obtained by solving the system of equations $\boldsymbol{u}_{\lambda}(\boldsymbol{\beta}) = \boldsymbol{0}_{d(k+1)}$, where*

$$\boldsymbol{u}_{\lambda}(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{u}_{\lambda}(\boldsymbol{\beta}, \boldsymbol{x}_{hi}), \tag{12}$$

$$\boldsymbol{u}_{\lambda}(\boldsymbol{\beta}, \boldsymbol{x}_{hi}) = \left[w_{hi}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{\widehat{\boldsymbol{y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right] \otimes \boldsymbol{x}_{hi}. \tag{13}$$

Let $\boldsymbol{U}_{\lambda}(\boldsymbol{\beta}, \boldsymbol{X})$ be a random variable generator of (13) associated with a generic random explanatory variable $\boldsymbol{X}$, with no stratum and cluster assignment. In what is to follow,

$$\boldsymbol{U}_{\lambda}(\boldsymbol{\beta}, \boldsymbol{x}_{hi}) = \left[w_{hi}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right] \otimes \boldsymbol{x}_{hi}$$

denotes $\boldsymbol{U}_{\lambda}(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}_{hi}$. An important property is that the estimating equation given in Theorem 2 is unbiased. This property is very important to consider these estimators similar to the maximum quasi weighted likelihood ones in the construction of the asymptotic properties, which was not the case for other distance based estimators such as the one proposed in Castilla et al (2018).

3.1 Asymptotic distribution and estimates of the design effect

The following results are generalized for the PLR model with complex sampling and random explanatory variables. Without any loss of generality it is assumed that $\widehat{\Pr}(\boldsymbol{X}=\boldsymbol{x}_{hi})=\frac{1}{n}$ is estimated from the sample of strata in order to get assymptotic results for the fixed explanatory variables (as a particular case of random explanatory variables).

Theorem 3 *Let $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$ the minimum quasi weighted DPD estimate of $\boldsymbol{\beta}$ in the PLR model (2) under a complex survey. Then we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{\beta}_0)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(\mathbf{0}_{d(k+1)},\boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}_0)\,\boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta}_0)\right),$$

where $\boldsymbol{\beta}_0$ is the true parameter value and

$$\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta})=\lim_{n\to\infty}\boldsymbol{\Omega}_{n,\lambda}(\boldsymbol{\beta})=\lim_{n\to\infty}\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{hi,\lambda}(\boldsymbol{\beta},\boldsymbol{\Sigma}_{hi}), \tag{14}$$

$$\boldsymbol{\Psi}_\lambda(\boldsymbol{\beta})=\lim_{n\to\infty}\boldsymbol{\Psi}_{n,\lambda}(\boldsymbol{\beta})=\lim_{n\to\infty}\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{\Psi}_{hi,\lambda}(\boldsymbol{\beta}), \tag{15}$$

where

$$\boldsymbol{\Omega}_{hi,\lambda}(\boldsymbol{\beta},\boldsymbol{\Sigma}_{hi})=w_{hi}^2\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Sigma}_{hi}\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T, \tag{16}$$

$$\boldsymbol{\Sigma}_{hi}=\mathrm{Var}[\widehat{\boldsymbol{Y}}_{hi}],$$

$$\boldsymbol{\Psi}_{hi,\lambda}(\boldsymbol{\beta})=\begin{cases} w_{hi}m_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T, & \lambda>0 \\ w_{hi}m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}^*(\boldsymbol{\beta}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T, & \lambda=0 \end{cases}. \tag{17}$$

Notice that the expression of $\boldsymbol{\Psi}_{\lambda=0}(\boldsymbol{\beta})$ is the same as the so called Fisher information matrix for multinomial sampling. In addition, if $\widehat{\boldsymbol{Y}}_{hi}$ has a multinomial sampling scheme $\boldsymbol{\Omega}_{\lambda=0}(\boldsymbol{\beta})=\boldsymbol{\Psi}_{\lambda=0}(\boldsymbol{\beta})$ and $\boldsymbol{\Omega}_{\lambda=0}^{-1}(\boldsymbol{\beta})\,\boldsymbol{\Psi}_{\lambda=0}(\boldsymbol{\beta})\,\boldsymbol{\Omega}_{\lambda=0}^{-1}(\boldsymbol{\beta})$ is the inverse of the Fisher information matrix.

Consistency is usually considered as a minimal requirement for an inference procedure. The following result is useful as a tool for estimating $\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta})$ and $\boldsymbol{\Psi}_\lambda(\boldsymbol{\beta})$ consistently plugging a consistent estimator into $\boldsymbol{\beta}$, and hence also $\boldsymbol{\Omega}_\lambda^{-1}(\boldsymbol{\beta})\,\boldsymbol{\Psi}_\lambda(\boldsymbol{\beta})\,\boldsymbol{\Omega}_\lambda^{-1}(\boldsymbol{\beta})$ again as a (double) plug-in estimator.

Corollary 4 *The following ones are (weak) consistent estimators as n goes to infinity:*

a) *$\widehat{\boldsymbol{\beta}}_{\lambda,Q}$ is a consistent estimator of the true regression coefficient $\boldsymbol{\beta}_0$.*

b) *$\boldsymbol{\Psi}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})=\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{\Psi}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$ is a consistent estimator of $\boldsymbol{\Psi}_\lambda(\boldsymbol{\beta}_0)$.*

c) *$\boldsymbol{\Omega}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q},\{\widehat{\boldsymbol{\Sigma}}_{hi}\}_{h=1,\ldots,H;i=1,\ldots,n_h})=\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q},\widehat{\boldsymbol{\Sigma}}_{hi})$ with*

$$\begin{aligned}\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q},\widehat{\boldsymbol{\Sigma}}_{hi})=w_{hi}^2\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\widehat{\boldsymbol{\Sigma}}_{hi}\\ \times\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T,\end{aligned} \tag{18}$$

is a consistent estimator of $\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}_0)$*, whenever* $\Sigma_{hi} = \mathrm{Var}[\widehat{\boldsymbol{Y}}_{hi}]$ *is consistently estimated through* $\widehat{\boldsymbol{\Sigma}}_{hi}$ *for all* $(h,i) \in \{1,...,H\} \times \{1,...,m_{hi}\}$.

The following two cases of $\boldsymbol{\Omega}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \{\widehat{\boldsymbol{\Sigma}}_{hi}\}_{h=1,...,H;i=1,...,n_h})$ are taken into account for two cases of $\boldsymbol{\Sigma}_{hi}$ for which there exists a consistent estimator:

- if $\widehat{\boldsymbol{Y}}_{hi}$ has (an ordinary) multinomial sampling scheme then

$$\boldsymbol{\Omega}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$$

  where $\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$ is $\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \widehat{\boldsymbol{\Sigma}}_{hi})$ given in (18) with

$$\widehat{\boldsymbol{\Sigma}}_{hi} = \boldsymbol{\Sigma}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}));$$

- if $\widehat{\boldsymbol{Y}}_{hi}$ has an overdispersed multinomial sampling scheme then

$$\boldsymbol{\Omega}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \{\widetilde{\nu}_{hi}\}_{h=1,...,H;i=1,...,n_h}) = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \widetilde{\nu}_{hi}),$$

  where $\widetilde{\nu}_{hi}$ is a consistent estimator of the overdispersion parameter ν_{hi} which is established later in Corollary 7, and $\boldsymbol{\Omega}_{n,\lambda}(\widetilde{\nu}_{hi}, \widehat{\boldsymbol{\beta}}_{\lambda,Q})$ is $\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \widehat{\boldsymbol{\Sigma}}_{hi})$ given in (18) with

$$\widehat{\boldsymbol{\Sigma}}_{hi} = \boldsymbol{\Sigma}_{hi}(\widetilde{\nu}_{hi}, \widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \widetilde{\nu}_{hi}m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}));$$

Remark 5 *If* $\widehat{\boldsymbol{Y}}_{hi}$ *has a multinomial sampling scheme, Theorem 3 can be similarly formulated taking the assumption that* $\widehat{\Pr}(\boldsymbol{X} = \boldsymbol{x}_{hi}) = \frac{1}{m}$*, where* $m = \sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$*, for each individual of all the clusters in the sample, and taking for the estimation equation* m *summands rather than* n*, i.e. plugging* $\sum_{j=1}^{m_{hi}} \boldsymbol{y}_{hij} = \widehat{\boldsymbol{y}}_{hi}$ *into (12) and considering the system of equations* $\boldsymbol{u}_\lambda(\boldsymbol{\beta}) = \boldsymbol{0}_{d(k+1)}$*, where*

$$\begin{aligned}
\boldsymbol{u}_\lambda(\boldsymbol{\beta}) &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\boldsymbol{v}_{j,\lambda}(\boldsymbol{\beta}, \boldsymbol{x}_{hi}),\\
\boldsymbol{v}_{j,\lambda}(\boldsymbol{\beta}, \boldsymbol{x}_{hi}) &= \left[w_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{\boldsymbol{y}_{hij} - \boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right] \otimes \boldsymbol{x}_{hi}.
\end{aligned}$$

Hence, it holds

$$\sqrt{m}(\widehat{\boldsymbol{\beta}}_{\lambda,Q} - \boldsymbol{\beta}_0) \xrightarrow[m\to\infty]{\mathcal{L}} \mathcal{N}\left(\boldsymbol{0}_{d(k+1)}, \boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}_0)\,\boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta}_0)\right),$$

with

$$\boldsymbol{\Omega}_{\lambda}\left(\boldsymbol{\beta}\right)=\lim_{m\to\infty}\begin{cases}\frac{1}{m}\sum_{h=1}^{H}\sum_{i=1}^{n_h}m_{hi}w_{hi}^{2}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right)\}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right)) & \lambda>0\\ \times\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right)\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^{T},\\ \frac{1}{m}\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}^{2}m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}^{*}\left(\boldsymbol{\beta}\right))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^{T}, & \lambda=0\end{cases}$$

$$\boldsymbol{\Psi}_{\lambda}\left(\boldsymbol{\beta}\right)=\lim_{m\to\infty}\begin{cases}\frac{1}{m}\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}m_{hi}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right)\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^{T}, & \lambda>0\\ \frac{1}{m}\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}^{*}\left(\boldsymbol{\beta}\right))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^{T}, & \lambda=0\end{cases}.$$

The formal proof is omitted, but the derivation of the expressions is almost the same considering $\boldsymbol{U}_{\lambda}(\boldsymbol{\beta},\boldsymbol{x}_{hi})=\sum_{j=1}^{m_{hi}}\boldsymbol{V}_{j,\lambda}(\boldsymbol{\beta},\boldsymbol{x}_{hi})$*, with* $\boldsymbol{V}_{j,\lambda}(\boldsymbol{\beta},\boldsymbol{x}_{hi})$ *i.i.d. random variables* $j=1,...,m_{hi}$*. This idea matches the philosophy of the asymptotic result developed in Castilla et al. (2019), where for* $H=1$ *and* $w_{1i}=1$*,* $i=1,...,m_{1i}$*.*

The following result is useful for any sample of polytomous logistic regression with complex sample design, more general in comparison with Corollary 4, since it is not necessary to get any consistent estimators for $\boldsymbol{\Sigma}_{hi}$.

Theorem 6 *The estimator* $\widehat{\boldsymbol{\Omega}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})=\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\widehat{\boldsymbol{\Omega}}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$*, with*

$$\begin{aligned}\widehat{\boldsymbol{\Omega}}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})&=\boldsymbol{U}_{\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q},\boldsymbol{x}_{hi})\boldsymbol{U}_{\lambda}^{T}(\widehat{\boldsymbol{\beta}}_{\lambda,Q},\boldsymbol{x}_{hi})\\&=\Big[w_{hi}^{2}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\{\widehat{\boldsymbol{Y}}_{hi}-m_{hi}\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\\&\times\{\widehat{\boldsymbol{Y}}_{hi}-m_{hi}\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}^{T}\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\Big]\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^{T},\end{aligned}$$

is consistent for $\boldsymbol{\Omega}_{\lambda}\left(\boldsymbol{\beta}\right)$ *as* n *goes to infinity.*

Corollary 7 *Let* $\widehat{\boldsymbol{Y}}_{hi}$ *be a random variable with overdisepersed multinomial sampling scheme with a common overdispersion parameter* ν *and* $m_{hi}=\overline{m}$*,*

$$\begin{aligned}\boldsymbol{\Sigma}_{hi}&=\boldsymbol{\Sigma}_{hi}\left(\nu,\boldsymbol{\beta}\right)=\nu\overline{m}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}\left(\boldsymbol{\beta}\right)),\\\nu&=1+\rho^{2}(\overline{m}-1),\end{aligned}$$

then, for ν *and* ρ^{2}*:*

a) *"robust and consistent estimators based on the estimating equation" are given respectively by*

$$\widetilde{\nu}_{n,\lambda}^{E}=\widetilde{\nu}_{n,\lambda}^{E}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})=\frac{1}{d(k+1)}\mathrm{trace}\left(\boldsymbol{\Omega}_{n,\lambda}^{-1}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\widehat{\boldsymbol{\Omega}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\right),\tag{19}$$

$$\widetilde{\rho}_{n,\lambda}^{2,E}=\frac{\widetilde{\nu}_{n,\lambda}^{E}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})-1}{\overline{m}-1},$$

where the matrices of interest are as follows, the one associated with multinomial sampling

$$\begin{aligned}\boldsymbol{\Omega}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}),\\ \boldsymbol{\Omega}_{hi,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \overline{m}w_{hi}^2\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\\ &\quad\times\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T,\end{aligned}$$

and Theorem 6 for overdispersed multinomial sampling.

b) *"robust and consistent estimators based on the method of moments" are given by*

$$\widetilde{\nu}_{n,\lambda}^{M} = \widetilde{\nu}_{n,\lambda}^{M}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \frac{1}{nd}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{d+1}\frac{(\widehat{Y}_{hij}-\overline{m}\pi_{hij}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))^2}{\overline{m}\pi_{hij}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})}, \tag{20}$$

$$\widetilde{\rho}_{n,\lambda}^{2,M} = \frac{\widetilde{\nu}_{n,\lambda}^{M}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})-1}{\overline{m}-1}.$$

The preceding results were established for $n=\sum_{h=1}^{H}n_h$ tending to infinity which implies in practice that H is fixed and there exists $\eta_h=\lim_{n\to\infty}\frac{n_h}{n}\in(0,1)$, $\sum_{h=1}^{H}\eta_h=1$. This means that it is approriate to consider in the place of Theorem 3, the following one, assuming that $\widehat{\Pr}(\boldsymbol{X}_h=\boldsymbol{x}_{hi})=\frac{1}{n_h}$.

Theorem 8 *Let $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$ the quasi MDPDE of $\boldsymbol{\beta}$ in the PLR model (2) under a complex survey. Then we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{\beta}_0)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(\mathbf{0}_{d(k+1)},\boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta}_0)\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta}_0)\right),$$

where $\boldsymbol{\beta}_0$ is the true parameter value and

$$\begin{aligned}\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}) &= \sum_{h=1}^{H}\eta_h\boldsymbol{\Omega}_\lambda^{(h)}(\boldsymbol{\beta}) = \sum_{h=1}^{H}\eta_h\lim_{n_h\to\infty}\frac{1}{n_h}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{i,\lambda}^{(h)}(\boldsymbol{\beta},\boldsymbol{\Sigma}_{hi}),\\ \boldsymbol{\Psi}_\lambda(\boldsymbol{\beta}) &= \sum_{h=1}^{H}\eta_h\boldsymbol{\Psi}_\lambda^{(h)}(\boldsymbol{\beta}) = \sum_{h=1}^{H}\eta_h\lim_{n_h\to\infty}\frac{1}{n_h}\sum_{i=1}^{n_h}\boldsymbol{\Psi}_{i,\lambda}^{(h)}(\boldsymbol{\beta}),\end{aligned}$$

where $\boldsymbol{\Omega}_{i,\lambda}^{(h)}(\boldsymbol{\beta},\boldsymbol{\Sigma}_{hi})$ and $\boldsymbol{\Psi}_{i,\lambda}^{(h)}(\boldsymbol{\beta})$ have the same expressions (not meaning) as $\boldsymbol{\Omega}_{hi,\lambda}(\boldsymbol{\beta},\boldsymbol{\Sigma}_{hi})$ and $\boldsymbol{\Psi}_{hi,\lambda}(\boldsymbol{\beta})$ respectively, given in (16), (17).

The idea of the previous theorem matches the philosophy of the asymptotic result developed in Castilla et al. (2018), where other family of divergences was considered and the derivations of the results were quite different. It is also appropriate to consider the following results in the place of Corollary 4, b and c.

Corollary 9 *The following ones are (weak) consistent estimators as n_h goes to infinity, for every $h=1,...,H$:*

a) $\boldsymbol{\Psi}_{\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \frac{1}{n_h}\sum_{i=1}^{n_h}\boldsymbol{\Psi}_{i,\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$ *is a consistent estimator of* $\boldsymbol{\Psi}_{\lambda}^{(h)}(\boldsymbol{\beta}_0)$*, for every* $h = 1, ..., H$*.*

b) $\boldsymbol{\Omega}_{\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \{\widehat{\boldsymbol{\Sigma}}_{hi}\}_{h=1,...,H:i=1,...,n_h}) = \frac{1}{n_h}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{i,\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \widehat{\boldsymbol{\Sigma}}_{hi})$ *is a consistent estimator of* $\boldsymbol{\Omega}_{\lambda}^{(h)}(\boldsymbol{\beta}_0)$*, for every* $h = 1, ..., H$*, whenever* $\Sigma_{hi} = \mathrm{Var}[\widehat{\boldsymbol{Y}}_{hi}]$ *is consistently estimated through* $\widehat{\boldsymbol{\Sigma}}_{hi}$ *for all* $(h, i) \in \{1, ..., H\} \times \{1, ..., m_{hi}\}$*.*

The following result is useful for any sample of polytomous logistic regression with complex sample design, more general in comparison with Corollary 9, since it is not neccessary to get any consistent estimators for $\boldsymbol{\Sigma}_{hi}$.

Theorem 10 *The estimator*

$$\begin{aligned}
\widehat{\boldsymbol{\Omega}}_{\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \frac{1}{n_h}\sum_{i=1}^{n_h}\widehat{\boldsymbol{\Omega}}_{i,\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}),\\
\widehat{\boldsymbol{\Omega}}_{i,\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \boldsymbol{U}_{\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{x}_{hi})\boldsymbol{U}_{\lambda}^{T}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{x}_{hi})\\
&= \Big[w_{hi}^2\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\{\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\\
&\quad \{\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}^{T}\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\Big] \otimes \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^{T},
\end{aligned}$$

is consistent of $\boldsymbol{\Omega}_{\lambda}^{(h)}(\boldsymbol{\beta})$*, as* n_h *goes to infinity, for every* $h = 1, ..., H$*.*

Corollary 11 *Let* $\widehat{\boldsymbol{Y}}_{hi}$ *be a random variable with overdisepersed multinomial sampling scheme with an overdispersion parameter* ν_h*, specific for each stratum, and* $m_{hi} = \overline{m}_h$*,*

$$\begin{aligned}
\boldsymbol{\Sigma}_{hi}(\boldsymbol{\beta}) &= \nu_h\overline{m}_h\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})),\\
\nu_h &= 1 + \rho_h^2(\overline{m}_h - 1),
\end{aligned}$$

then, for ν_h *and* ρ_h^2*:*

a) *"robust and consistent estimators based on the estimating equation" are given respectively by*

$$\begin{aligned}
\widetilde{\nu}_{h,\lambda}^{E}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \frac{1}{d(k+1)}\mathrm{trace}\left(\boldsymbol{\Omega}_{\lambda}^{(h),-1}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\widehat{\boldsymbol{\Omega}}_{\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\right),\\
\widetilde{\rho}_{h,\lambda}^{2,E} &= \frac{\widetilde{\nu}_{h,\lambda}^{E}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) - 1}{\overline{m}_h - 1},
\end{aligned}$$

where the matrices of interest are the one associated with multinomial sampling for the h-th stratum,

$$\begin{aligned}
\boldsymbol{\Omega}_{\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \frac{1}{n_h}\sum_{i=1}^{n_h}\boldsymbol{\Omega}_{i,\lambda}^{(h)}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}),\\
\boldsymbol{\Omega}_{i,\lambda}^{(h)}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{hi}) &= \overline{m}_h w_{hi}^2\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\\
&\quad \times \mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}\widehat{\boldsymbol{\beta}}_{\lambda,Q})) \otimes \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^{T}
\end{aligned}$$

and Theorem 10 for overdispersed multinomial sampling.

b) *"robust and consistent estimators based on the method of moments" are given by*

$$\widetilde{\nu}_{h,\lambda}^{M}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \frac{1}{n_h d}\sum_{i=1}^{n_h}\sum_{j=1}^{d+1}\frac{(\widehat{Y}_{hij} - \overline{m}_h\pi_{hij}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))^2}{\overline{m}_h\pi_{hij}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})},$$

$$\widetilde{\rho}_{h,\lambda}^{2,M} = \frac{\widetilde{\nu}_{h,\lambda}^{M}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) - 1}{\overline{m}_h - 1}.$$

4 Testing linear hypotheses for the PLR coefficients with complex survey design

Based on the asymptotic distribution of the minimum quasi weighted DPD estimator, $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$, presented in Theorem 8, we now introduce and study a family of Wald-type test statistics for testing

$$H_0 : \boldsymbol{M}^T\boldsymbol{\beta} = \boldsymbol{l} \text{ against } H_0 : \boldsymbol{M}^T\boldsymbol{\beta} \neq \boldsymbol{l} \tag{21}$$

where $\boldsymbol{M}$ is a $d(k+1)\times r$ full row-rank matrix with $r \leq d(k+1)$ and the right-hand-side $r-$vector consists of constants that in many situations are $\boldsymbol{l} = \boldsymbol{0}_r$. For solving the problem of testing given in (21) we define a family of Wald-type test statistics based on quasi MDPDE as follows.

Definition 12 *Let $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$ the minimum quasi weighted DPD estimator of $\boldsymbol{\beta}$ in the PLR model (2) under a complex survey and we denote*

$$\begin{aligned}\widehat{\mathbf{Q}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \frac{1}{n}\boldsymbol{\Psi}_{n,\lambda}^{-1}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\widehat{\boldsymbol{\Omega}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\boldsymbol{\Psi}_{n,\lambda}^{-1}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\\ &= \left(n\boldsymbol{\Psi}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\right)^{-1} n\widehat{\boldsymbol{\Omega}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\left(n\boldsymbol{\Psi}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\right)^{-1},\end{aligned}$$

where

$$\begin{aligned}n\widehat{\boldsymbol{\Omega}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\Big[&w_{hi}^2\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\{\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\\ &\times\{\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}^T\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T,\end{aligned}$$

$$n\boldsymbol{\Psi}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \begin{cases}\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}m_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T, & \lambda > 0\\ \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}m_{hi}\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}^*(\boldsymbol{\beta}))\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T, & \lambda = 0\end{cases}.$$

The family of Wald-type test statistics for testing the null hypothesis given in (21) is given by

$$W_n(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = (\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_{\lambda,Q} - \boldsymbol{l})^T\left(\boldsymbol{M}^T\widehat{\mathbf{Q}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\boldsymbol{M}\right)^{-1}(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_{\lambda,Q} - \boldsymbol{l}) \tag{22}$$

For $\lambda = 0$, $\widehat{\boldsymbol{\beta}}_{\lambda=0,Q}$, is the maximum quasi weighted likelihood estimator of $\boldsymbol{\beta}$, with estimating equations given in (9). It is not difficult to see that $\mathbf{Q}_{\lambda=0}(\widehat{\boldsymbol{\beta}}_{\lambda=0,Q})$ is the Fisher information matrix and $W_n(\widehat{\boldsymbol{\beta}}_{\lambda=0,Q})$ will be the classical Wald test statistic.

Theorem 13 *Under the null hypothesis given in (21) the asymptotic distribution of the Wald-type test statistics $W_n(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$, defined in (22), is chi-square with r degrees of freedom.*

The proof is immediate using the asymptotic distribution of the minimum quasi weighted DPD estimator, presented in Theorem 8, and taking into account the consistence of the matrix $\mathbf{Q}_\lambda\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}\right)$ presented in Corollary 9. Based on the previous theorem the null hypothesis given in (21) will be rejected if

$$W_n\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}\right) > \chi^2_{r,\alpha}. \tag{23}$$

Next we present a result in order to give an approximation of the power function for the test statisitcs given in (23). Let $\boldsymbol{\beta}^* \in \Theta$ be such that $\boldsymbol{M}^T\boldsymbol{\beta}^* \neq \boldsymbol{c}$, i.e., $\boldsymbol{\beta}^*$ does not belong to the null parameter space. Let us denote

$$l_{\boldsymbol{\beta}_1}(\boldsymbol{\beta}_2) = \left(\boldsymbol{M}^T\boldsymbol{\beta}_1 - \boldsymbol{l}\right)^T \left(\boldsymbol{M}^T\widehat{\mathbf{Q}}_{n,\lambda}\left(\boldsymbol{\beta}_2\right)\boldsymbol{M}\right)^{-1}\left(\boldsymbol{M}^T\boldsymbol{\beta}_1 - \boldsymbol{l}\right).$$

Then we have the following result.

Theorem 14 *Let $\boldsymbol{\beta}^* \in \Theta$ with $\boldsymbol{M}^T\boldsymbol{\beta}^* \neq \boldsymbol{l}$ the true value of the parameter so that $\widehat{\boldsymbol{\beta}}_{\lambda,Q} \xrightarrow[n\to\infty]{P} \boldsymbol{\beta}^*$. The power function of the test statistic given in (23), at $\boldsymbol{\beta}^*$, is given by*

$$Po_{W_n(\widehat{\boldsymbol{\beta}}_{\lambda,Q})}\left(\boldsymbol{\beta}^*\right) = 1 - \phi_n\left(\frac{1}{\sigma\left(\boldsymbol{\beta}^*\right)}\left(\frac{\chi^2_{r,\alpha}}{\sqrt{n}} - \sqrt{n}l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)\right) \tag{24}$$

where $\phi_n(x)$ tends uniformly to the standard normal distribution $\phi(x)$ and $\sigma\left(\boldsymbol{\beta}^*\right)$ is given by

$$\sigma\left(\boldsymbol{\beta}^*\right)^2 = \left.\frac{\partial l_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \widehat{\mathbf{Q}}_{n,\lambda}\left(\boldsymbol{\beta}^*\right)\left.\frac{\partial l_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

It is clear that

$$\lim_{n\to\infty} Po_{W_n(\widehat{\boldsymbol{\beta}}_{\lambda,Q})}\left(\boldsymbol{\beta}^*\right) = 1$$

for all $\alpha \in (0,1)$. Therefore, the Wald-type test statistics are consistent in the sense of Fraser.

Remark 15 *Based on the previous theorem we can obtain the sample size necessary to get a fix power $Po_{W_n(\widehat{\boldsymbol{\beta}}_{\lambda,Q})}\left(\boldsymbol{\beta}^*\right) = \pi_0$. By (24) we must solve the equation*

$$1 - \pi_0 = \phi\left(\frac{1}{\sigma\left(\boldsymbol{\beta}^*\right)}\left(\frac{\chi^2_{r,\alpha}}{\sqrt{n}} - \sqrt{n}l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)\right)$$

and we get that $n = [n^*] + 1$ with

$$n^* = \frac{A + B + \sqrt{A(A+2B)}}{2l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)^2}$$

being $A = \sigma^2\left(\boldsymbol{\beta}^*\right)\left(\phi^{-1}\left(1-\pi_0\right)\right)^2$ and $B = \frac{1}{2}\chi^2_{r,\alpha}l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)$.

We may also find an approximation of the power of the Wald-type tests given in (23) at an alternative close to the null hypothesis. Let $\boldsymbol{\beta}_n$ such that $\boldsymbol{M}^T\boldsymbol{\beta_n} \neq \boldsymbol{l}$ be a given alternative and let $\boldsymbol{\beta}_0$ be the element such that $\boldsymbol{M}^T\boldsymbol{\beta_0} = \boldsymbol{l}$ closest to $\boldsymbol{\beta}_n$ in the Euclidean distance sense. A first possibility to introduce contiguous alternative hypotheses is to consider a fixed $\boldsymbol{d} \in \mathbb{R}^{d(k+1)}$ and to permit $\boldsymbol{\beta}_n$ moving towards $\boldsymbol{\beta}_0$ as n increases in the following way

$$H_{1,n} : \boldsymbol{\beta} = \boldsymbol{\beta}_n, \quad \text{where} \quad \boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{d}. \tag{25}$$

A second approach is to relax the condition $\boldsymbol{M}^T\boldsymbol{\beta_0} = \boldsymbol{l}$ defining the null hypothesis. Let $\boldsymbol{\delta} \in \mathbb{R}^r$ and consider the following sequence, $\boldsymbol{\beta}_n$, of parameters moving towards $\boldsymbol{\beta}_0$ according to

$$H^*_{1,n} : \boldsymbol{\beta} = \boldsymbol{\beta}_n, \quad \text{where} \quad \boldsymbol{M}^T\boldsymbol{\beta_n} - \boldsymbol{l} = n^{-1/2}\boldsymbol{\delta}. \tag{26}$$

Note that

$$\boldsymbol{M}^T\boldsymbol{\beta_n} - l = \boldsymbol{M^T\beta_0} + \boldsymbol{M^T}\, n^{-1/2}\boldsymbol{d} - \boldsymbol{l} = M^{\boldsymbol{T}} n^{-1/2}\boldsymbol{d}. \tag{27}$$

Then the equivalence between the two hypotheses is given by $M^{\boldsymbol{T}}\boldsymbol{d} = \boldsymbol{\delta}$.

If we denote by $\chi^2_r(\Delta)$ the non central chi-square distribution with r degrees of freedom and noncentrality parameter Δ, we can state the following theorem.

Theorem 16 *We have,*

i) $W_n\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}\right) \underset{n\longrightarrow\infty}{\overset{\mathcal{L}}{\longrightarrow}} \chi^2_r\left(\Delta_1\right)$*under* $H_{1,n}$ *given in (25), with*

$$\Delta_1 = \boldsymbol{d}^T\boldsymbol{M}\left[M^{\boldsymbol{T}}\widehat{\mathbf{Q}}_{\boldsymbol{n},\boldsymbol{\lambda}}\left(\boldsymbol{\beta_0}\right)\boldsymbol{M}\right]^{-1} M^{\boldsymbol{T}}\boldsymbol{d}.$$

ii) $W_n\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}\right) \underset{n\longrightarrow\infty}{\overset{\mathcal{L}}{\longrightarrow}} \chi^2_r\left(\Delta_2\right)$ *under* $H^*_{1,n}$ *given in (26), with*

$$\Delta_2 = \boldsymbol{\delta}^T\left(M^{\boldsymbol{T}}\widehat{\mathbf{Q}}_{\boldsymbol{n},\boldsymbol{\lambda}}\left(\boldsymbol{\beta_0}\right)\boldsymbol{M}\right)^{-1}\boldsymbol{\delta}.$$

5 Influence Function

Influence function is a classical tool to measure robustness of an estimator (Hampel et al., 1968). However, the present set-up of complex survey is not as simple as the iid set-up; in fact the observations within a cluster of a stratum are iid but the observations in different cluster and stratum are independent non-homogeneous. So we need to modify the definition of the influence function accordingly. Recently, Ghosh and Basu (2013, 2016, 2018) have discussed the extended definition of the influence function for the independent but non-homogeneous observations; we will extend their approach to define the influence function in the present case of PLR model under complex design.

5.1 Influence Function of the minimum quasi weighted DPD estimator

We first need to define the statistical functional corresponding to the minimum quasi weighted DPD estimator as the minimizer of the DPD between the true and model densities. Assume the set-up and notations of Section 3 with the DPD kernel being given by $d^*_\lambda\left(g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right)$ for individual densities $g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ and $f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij})$. Then, following Ghosh and Basu (2013), the minimum quasi weighted DPD estimator functional is to be defined by the minimizer of the total weighted DPD measure given by

$$\begin{aligned} d_\lambda(g, f_{\boldsymbol{\beta}}, w) &= \sum_{h=1}^{H}\sum_{i=1}^{n_h} d^*_\lambda\left(g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}), f_{\boldsymbol{\beta}}(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})\right) \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}\left(m_{hi}\boldsymbol{\pi}_{hi}^{\lambda,T}(\boldsymbol{\beta})\,\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}) - \frac{\lambda+1}{\lambda}\int \boldsymbol{\pi}_{hi}^{\lambda,T}(\boldsymbol{\beta})\,\boldsymbol{y}_{hi}dG(\boldsymbol{y}_{hi}|\boldsymbol{x}_{hi})\right), \quad \text{for } \lambda > 0. \end{aligned} \tag{28}$$

Definition 17 *We consider the PLR model with complex survey defined in (2). The minimum quasi weighted DPD estimator functional, $\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G})$, of $\boldsymbol{\beta}$ at $\boldsymbol{G} = (G(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi}),\ h = 1,\ldots,H, i = 1,\ldots,n_h)$ is defined as*

$$\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}) = \arg\min_{\boldsymbol{\beta}\in\Theta} d_\lambda(g, f_{\boldsymbol{\beta}}, w),$$

where $d_\lambda(g, f_{\boldsymbol{\beta}}, w)$ is as defined above in (28).

Note that, by the property of the DPD measure, it is immediate that the minimum quasi weighted DPD estimator functional $\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G})$ is Fisher consistent at the assumed PLR model (2), i.e., $\boldsymbol{T}_{\lambda,Q}(\boldsymbol{\pi}(\boldsymbol{\beta})) = \boldsymbol{\beta}$ for all parameter values $\boldsymbol{\beta}$. Also, following the calculations for Theorem 2, one can see that the minimum quasi weighted DPD estimator functional $\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G})$ can also be derived as a solution to the estimating equations

$$\boldsymbol{u}^*_\lambda(\boldsymbol{g}, \boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \boldsymbol{u}^*_{\lambda,hi}(g_{hi}, \boldsymbol{\beta}) = \boldsymbol{0}_{d(k+1)}, \tag{29}$$

where $g_{hi} = g(\boldsymbol{y}_{hij}|\boldsymbol{x}_{hi})$ and

$$\boldsymbol{u}^*_{\lambda,hi}(g_{hi}, \boldsymbol{\beta}) = \left[w_{hi}m_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{g_{hi} - \boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right] \otimes \boldsymbol{x}_{hi}. \tag{30}$$

Note that, these estimating equations are unbiased at the model probability $\boldsymbol{g} = \boldsymbol{\pi}(\boldsymbol{\beta})$.

Now, in order to define the influence function of the minimum DPD functional $\boldsymbol{T}_{\lambda,Q}(\boldsymbol{g})$, we note that the functional itself depends on the sample sizes and cluster weights and so it IF will have the same dependence in analogue to the non-homogeneous case of Ghosh and Basu (2013, 2016, 2018). Also note that, the contamination can be in any one particular cluster within one stratum or simultaneously in many of them (or all). For simplicity, let us first assume that the contamination is only in one

cluster probability $g_{h_0 i_0}$ for some fix h_0 and i_0. Consider the contaminated probability vector $g_{h_0 i_0, \epsilon} = (1-\epsilon) g_{h_0 i_0} + \epsilon \delta_{\boldsymbol{t}}$, where ϵ is the contamination proportion and $\delta_{\boldsymbol{t}}$ is the degenerate probability at the outlier point $\boldsymbol{t} = (t_1, \ldots, t_{d+1})^T \in \{0,1\}^{d+1}$ with $\sum_{s=1}^{d+1} t_s = 1$. Denote the corresponding contaminated full probability vector as $\boldsymbol{g}_\epsilon$ which is same as $\boldsymbol{g}$ except $g_{h_0 i_0}$ being replaced by $g_{h_0 i_0, \epsilon}$ and let the corresponding contaminated distribution vector be $\boldsymbol{G}_\epsilon$. Then, the corresponding influence function is defined as

$$\mathcal{IF}(\boldsymbol{t}, \boldsymbol{T}_{\lambda,Q}, \boldsymbol{g}) = \lim_{\epsilon \to 0} \frac{\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}_\epsilon) - \boldsymbol{T}_{\lambda,Q}(\boldsymbol{G})}{\epsilon} = \frac{\partial}{\partial \epsilon} \boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}_\epsilon) \Big|_{\epsilon=0}.$$

In order to calculate this influence function, we start with the estimating equation for $\boldsymbol{T}_{\lambda,Q}$. Note that, $\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}_\epsilon)$ satisfies the equations

$$\boldsymbol{u}^*_\lambda \left(\boldsymbol{g}_\epsilon, \boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}_\epsilon) \right) = \boldsymbol{0}_{d(k+1)}.$$

Now, differentiating it with respect to ϵ art $\epsilon = 0$ and simplifying, we get the required influence function as given by

$$\mathcal{IF}(\boldsymbol{t}, \boldsymbol{T}_{\lambda,Q}, \boldsymbol{g}) = \boldsymbol{\Psi}^*_{n,\lambda}(\boldsymbol{g}, \boldsymbol{\beta})^{-1} \frac{1}{n} \left[\boldsymbol{u}^*_{\lambda, h_0 i_0} \left(\delta_{\boldsymbol{t}}, \boldsymbol{\beta} \right) - \boldsymbol{u}^*_{\lambda, h_0 i_0} \left(g_{h_0 i_0}, \boldsymbol{\beta} \right) \right], \tag{31}$$

where

$$\boldsymbol{\Psi}^*_{n,\lambda}(\boldsymbol{g}, \boldsymbol{\beta}) = -\frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{u}^*_{\lambda, hi} \left(g_{hi}, \boldsymbol{\beta} \right). \tag{32}$$

Note that, at the model probability $\boldsymbol{g} = \boldsymbol{\pi}(\boldsymbol{\beta})$, we have $\boldsymbol{u}^*_{\lambda, h_0 i_0} \left(\boldsymbol{\pi}_{h_0 i_0}(\boldsymbol{\beta}), \boldsymbol{\beta} \right) = \boldsymbol{0}$ and $\boldsymbol{\Psi}^*_{n,\lambda}(\boldsymbol{\pi}(\boldsymbol{\beta}), \boldsymbol{\beta}) = \boldsymbol{\Psi}_{n,\lambda}(\boldsymbol{\beta})$, as defined in Theorem 3. Hence, the influence function of the proposed minimum quasi weighted DPD estimator at the model probability simplifies to

$$\mathcal{IF}(\boldsymbol{t}, \boldsymbol{T}_{\lambda,Q}, \boldsymbol{\pi}(\boldsymbol{\beta})) = \boldsymbol{\Psi}_{n,\lambda}(\boldsymbol{\beta})^{-1} \frac{1}{n} \boldsymbol{u}^*_{\lambda, h_0 i_0} \left(\delta_{\boldsymbol{t}}, \boldsymbol{\beta} \right). \tag{33}$$

Clearly, by the assumed form of the model probability $\boldsymbol{\pi}(\boldsymbol{\beta})$ this influence function is bounded for all $\lambda > 0$ but unbounded at $\lambda = 0$. So, the proposed minimum quasi weighted DPD estimators are expected to be robust for any $\lambda > 0$ but the corresponding maximum quasi weighted likelihood estimator (at $\lambda = 0$) is non-robust against data contamination.

Similarly, one can show that, if there is contamination in some of the clusters within some stratum indexed by $h, i \in \Gamma \subseteq \{h = 1, \ldots, H; i = 1, \ldots, n_h\}$ at the contamination points $\boldsymbol{t}_{hi}$, the corresponding influence function of the proposed minimum quasi weighted DPD estimator at $\boldsymbol{g}$ and the model $\pi(\boldsymbol{\beta})$ are respectively given by

$$\mathcal{IF}((\boldsymbol{t} : \{h,i\} \in \Gamma), \boldsymbol{T}_{\lambda,Q}, \boldsymbol{g}) = \boldsymbol{\Psi}^*_{n,\lambda}(\boldsymbol{g}, \boldsymbol{\beta})^{-1} \frac{1}{n} \sum_{\{h,i\} \in \Gamma} \left[\boldsymbol{u}^*_{\lambda, hi} \left(\delta_{\boldsymbol{t}_{hi}}, \boldsymbol{\beta} \right) - \boldsymbol{u}^*_{\lambda, hi} \left(g_{hi}, \boldsymbol{\beta} \right) \right], \tag{34}$$

and

$$\mathcal{IF}((\boldsymbol{t} : \{h,i\} \in \Gamma), \boldsymbol{T}_{\lambda,Q}, \boldsymbol{\pi}(\boldsymbol{\beta})) = \boldsymbol{\Psi}_{n,\lambda}(\boldsymbol{\beta})^{-1} \frac{1}{n} \sum_{\{h,i\} \in \Gamma} \boldsymbol{u}^*_{\lambda, hi} \left(\delta_{\boldsymbol{t}_{hi}}, \boldsymbol{\beta} \right). \tag{35}$$

The boundedness and robustness implications for these influence functions are exactly the same as before.

5.2 Influence Function of the Wald-type Tests in PLRM

Let us now study the robustness of proposed Wald-type test through the influence function of the corresponding test statistics in (22). Considering the notation and set-up of Section 2 and 3, and define

$$\mathbf{Q}_\lambda(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta})\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta})\boldsymbol{\Psi}_\lambda^{-1}(\boldsymbol{\beta}).$$

Then the Wald-type test functional $W_\lambda(\boldsymbol{G})$ corresponding to (22) for testing the null hypothesis given in (21) at the true distribution vector $\boldsymbol{G}$ is defined as

$$W_\lambda(\boldsymbol{G}) = (\boldsymbol{M}^T\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}) - \boldsymbol{l})^T \left(\boldsymbol{M}^T\mathbf{Q}_\lambda(\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}))\boldsymbol{M}\right)^{-1} (\boldsymbol{M}^T\boldsymbol{T}_{\lambda,Q}(\boldsymbol{G}) - \boldsymbol{l}), \tag{36}$$

where $\boldsymbol{T}_{\lambda,Q}$ is the minimum quasi weighted DPD estimator functional as defined in the previous subsection. Also, let $\boldsymbol{\beta}_0$ denote the true null parameter value for the hypothesis in (21).

Let us now derive the influence function of the test functional W_λ. As before, here also the contamination can be any particular cluster and strata (a given h_0, i_0) combination or in many (or all) of them. The influence function of general Wald-type tests under such non-homogeneous set-up has been extensively studied in Basu et al. (2018). Here, we follow the general theory of Basu et al. (2018) to conclude that the first order influence functions of W_λ, defined as the first order derivative of its value at the contaminated distribution with respect to ϵ at $\epsilon = 0$, in either case of contamination become identically zero at the null distribution $\boldsymbol{g} = \boldsymbol{\pi}(\boldsymbol{\beta}_0)$. Therefore, the first order influence function is not informative in this case of Wald-type tests, and we need to investigate the second order influence function of W_λ.

The second order influence function $IF^{(2)}$, which measures the second order approximation to the asymptotic bias due to infinitesimal contamination, is defined as the second order derivative of the value of W_λ at the contaminated distribution with respect to the contamination proportion ϵ at $\epsilon = 0$. Again, following Basu et al. (2018), we can derive these second order influence functions of the Wald-type tests in either case of contaminations; at the null distribution $\boldsymbol{g} = \boldsymbol{\pi}(\boldsymbol{\beta}_0)$, they simplifies to

$$\begin{aligned}
&\mathcal{IF}^{(2)}(\boldsymbol{t}, W_\lambda, \boldsymbol{\pi}(\boldsymbol{\beta}_0)) = 2\mathcal{IF}(\boldsymbol{t}, \boldsymbol{T}_{\lambda,Q}, \boldsymbol{\pi}(\boldsymbol{\beta}_0))^T\boldsymbol{M}\left(\boldsymbol{M}^T\mathbf{Q}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\mathcal{IF}(\boldsymbol{t}, \boldsymbol{T}_{\lambda,Q}, \boldsymbol{\pi}(\boldsymbol{\beta}_0)),\\
&\mathcal{IF}^{(2)}((\boldsymbol{t} : \{h,i\} \in \Gamma), W_\lambda, \boldsymbol{\pi}(\boldsymbol{\beta}_0))\\
&= 2\mathcal{IF}((\boldsymbol{t} : \{h,i\} \in \Gamma), \boldsymbol{T}_{\lambda,Q}, \boldsymbol{\pi}(\boldsymbol{\beta}_0))^T\boldsymbol{M}\left(\boldsymbol{M}^T\mathbf{Q}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\mathcal{IF}((\boldsymbol{t} : \{h,i\} \in \Gamma), \boldsymbol{T}_{\lambda,Q}, \boldsymbol{\pi}(\boldsymbol{\beta}_0)),
\end{aligned}$$

for the two types of contamination as before in the case of estimator. Note that, the second order influence functions of the proposed Wald-type tests are a quadratic function of the corresponding influence functions of the minimum quasi weighted DPD estimator for any type of contamination. Therefore, the boundedness of the influence functions of minimum quasi weighted DPD estimator at $\lambda > 0$ also indicates the boundedness of the influence functions of the Wald-type test functional W_λ indicating their robustness against contamination in any cluster or stratum of the sample data.

6 Illustrative Example: BMI data set

Let us consider an illustrative real-life dataset on BMI which was previously studied in Castilla et al. (2018). This data set, obtained from CANSIM Canada's database and presented in Table 1, shows the body mass indexes of population in Canada in the year 1994. Each person in the sample is divided by their body mass index category under the international standard: acceptable weight, overweight or obese. The data set consists on a stratified sample design with clusters nested on them, with the strata being three different age groups (20–34 years, 35–44 years and 45–64 years) and the genders (male or female) as the clusters. The qualitative explanatory variables are valid to distinguish the clusters within the strata. They are given by $\boldsymbol{x}_{h1}^T = (1, 0)$, and $\boldsymbol{x}_{h2}^T = (0, 1)$, $h = 1, \ldots, 5$ for men and women, respectively.

Table 1: Body mass index (BMI) data set

		Body mass index (BMI)		
Age group	Sex	Acceptable weight (18.5-24.9)	Overweight (25.0-29.9)	Obese (30 or higher)
20-34 years	Men	5438	4790	1470
	Women	4910	2878	802
35-44 years	Men	2458	3437	1319
	Women	3100	1494	1313
45-64 years	Men*	1968	3290	1412
	Women*	1710	1481	1078

To illustrate the robustness of minimum quasi weighted DPD estimators, we contaminate the BMI data by permuting the categories overweight and obese in the Men with age range 45–64 years. After obtaining the correpsonding minimum quasi weighted DPD estimators estimates, we compute the mean absolute standardized deviations ($mssd$) between the estimated parameters and corresponding estimated probabilities obtained for the modified and original data, as given by

$$masd(\widehat{\boldsymbol{\beta}}_\lambda^*, \widehat{\boldsymbol{\beta}}_\lambda) = \frac{1}{4}\sum_{r=1}^{2}\sum_{s=1}^{2}\left|\frac{\widehat{\beta}_{\lambda,rs}^* - \widehat{\beta}_{\lambda,rs}}{\widehat{\beta}_{\lambda,rs}}\right| \quad \text{and} \quad masd(\widehat{\boldsymbol{\pi}}_\lambda^*, \widehat{\boldsymbol{\pi}}_\lambda) = \frac{1}{6}\sum_{r=1}^{3}\sum_{s=1}^{2}\left|\frac{\pi_{rs}(\widehat{\boldsymbol{\beta}}_\lambda^*) - \pi_{rs}(\widehat{\boldsymbol{\beta}}_\lambda)}{\pi_{rs}(\widehat{\boldsymbol{\beta}}_\lambda)}\right|,$$

where, with superscript * we denote the contaminated case. Assuming common overdispersion parameter, absolute standardized deviation (asd) of the two versions of the intra-cluster correlation estimator are also computed as in Corollary 7. Their values, as presented in Table 2, clearly show that the minimum quasi weighted DPD estimators become more robust as λ increases. Similar results are obtained when permuting the categories for Women instead of Men, as can also be seen in Table 2.

Table 2: Mean standardized deviations under contamination for BMI data

	λ					
	0	0.2	0.4	0.6	0.8	1
contamination of Men (45–64 years)						
$masd(\widehat{\boldsymbol{\beta}}^*_\lambda,\widehat{\boldsymbol{\beta}}_\lambda)$	0.24396	0.23057	0.21731	0.20441	0.19187	0.17969
$masd(\widehat{\boldsymbol{\pi}}^*_\lambda,\widehat{\boldsymbol{\pi}}_\lambda)$	0.10170	0.09700	0.09220	0.08750	0.08280	0.07810
$asd(\widehat{\rho}^{2*,E}_\lambda,\widehat{\rho}^{2,E}_\lambda)$	0.64785	0.58974	0.52955	0.46868	0.40858	0.35044
$asd(\widehat{\rho}^{2*,M}_\lambda,\widehat{\rho}^{2,M}_\lambda)$	1.57103	1.41605	1.25700	1.09724	0.94068	0.79037
contamination of Women (45–64 years)						
$masd(\widehat{\boldsymbol{\beta}}^*_\lambda,\widehat{\boldsymbol{\beta}}_\lambda)$	0.10516	0.09484	0.08533	0.07665	0.0687	0.06148
$masd(\widehat{\boldsymbol{\pi}}^*_\lambda,\widehat{\boldsymbol{\pi}}_\lambda)$	0.03250	0.03030	0.02810	0.02600	0.0240	0.02210
$asd(\widehat{\rho}^{2*,E}_\lambda,\widehat{\rho}^{2,E}_\lambda)$	0.07038	0.05132	0.03358	0.01744	0.00321	0.00893
$asd(\widehat{\rho}^{2*,M}_\lambda,\widehat{\rho}^{2,M}_\lambda)$	0.03358	0.01169	0.05244	0.08819	0.11821	0.14206

7 Monte Carlo Simulation Study

7.1 Simulation Scheme

We develop a complex design extension of the simulation scheme previously studied by Castilla et al. (2019), where a simple sample design was considered. H strata consisting of $n_h = n$ clusters having $m_{hi} = m$ units each, are taken. Three overdispersed multinomial distributions: the random-clumped (RC), the m-inflated (m-I) and the Dirichlet Multinomial (DM) distributions (Alonso et al., 2017) having the same parameters $\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}_0)$ and ρ, are then considered for $\widehat{\boldsymbol{Y}}_{hi}$; which are characterized by

$$\boldsymbol{E}[\widehat{\boldsymbol{Y}}_{hi}] = m\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}_0) \quad \text{and} \quad \boldsymbol{V}[\widehat{\boldsymbol{Y}}_{hi}] = \nu_m m\boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}_0)),$$
$$\nu_m = 1 + \rho^2(m-1), \qquad i = 1,\ldots,n;\ h = 1,2.$$

As in Castilla et al. (2019), we further assume that the outcome nominal variable $\boldsymbol{Y}$ has $d+1=3$ categories depending on $k=2$ explanatory variables through the PLR model probabilities

$$\pi_{hir}(\boldsymbol{\beta}) = \begin{cases} \dfrac{\exp\{\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_r\}}{1+\sum_{l=1}^d \exp\{\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_l\}}, & r=1,2 \\ \dfrac{1}{1+\sum_{l=1}^d \exp\{\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_l\}}, & r=3 \end{cases},$$

with $\boldsymbol{\beta} = (\beta_{01},\beta_{11},\beta_{21},\beta_{02},\beta_{12},\beta_{22})^T = (0,-0.9,0.1,0.6,-1.2,0.8)^T$ and $\boldsymbol{x}_{hi} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$ for all $i = 1,\ldots,n$, $h=1,\ldots,H$.

Considering different values of the intra-cluster correlation parameter (ρ^2), the number of clusters in each stratum (n) and the clusters sizes (m), we simulate data from different scenarios.

Scenario 1: $H = 2$, $n \in \{10i\}_{i=4}^{15}$, $m = 21$, $\rho^2 = 0.25$, RC distribution.

Scenario 1b: $H = 2$, $n \in \{10i\}_{i=4}^{15}$, $m = 21$, $\rho^2 = 0.50$, RC distribution.

Scenario 2: $H = 2$, $n = 60$, $m \in \{10i\}_{i=2}^{12}$, $\rho^2 = 0.25$, RC m-I and DM distributions.

Scenario 3: $H = 2$, $n = 60$, $m = 21$, $\rho^2 \in \{0.1i\}_{i=0}^{9}$, RC, m-I and DM distributions.

In order to study the robustness issue, these simulations are repeated under contaminated data having 7% outliers. These outliers are generated by permuting the elements of the outcome variable, such that categories 1, 2, 3 are classified as categories 3, 1, 2 for the outlying observations.

7.2 Performance of minimum quasi weighted DPD estimators and ρ^2 estimates

For the above scenarios, we compute the minimum quasi weighted DPD estimator of $\boldsymbol{\beta}$, for different tuning parameters $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8\}$, and the corresponding estimate of ρ^2, both for the methods of moments and the estimating equations. The root of the mean square error (RMSE) are then computed based on 1000 replications (Figures 5–8).

While classical maximum quasi weighted likelihood estimator presents the best behaviour under pure data, minimum quasi weighted DPD estimators with $\lambda > 0$ are much more robust. In particular, as λ increases, the change on their behaviour is accentuated. This is independent to the sample size (Figures 5 and 6), the intra-cluster correlation and the distribution (Figure 7) considered.

Estimator of ρ^2 by the method of moments seems less precise than the method of estimating equations, with independence of the tuning parameter chosen (Figures 5 and 6). Best estimators of ρ^2 by the estimating equations are obtained from minimum quasi weighted DPD estimators with $\lambda > 0$, both for pure and contaminated data and for any of the distributions considered (Figure 7). Error of the estimators of ρ^2 tends to be smaller with the DM and RC distributions in comparison with the mI distribution, as can be seen in Figure 1, where density plots based on $1,000$ replications are shown for $\rho^2 = 0.5, n = 60, m = 21$ and tuning parameter $\lambda = 0.4$. These results on the design effect parameter are consistent with the previous work of Alonso et al. (2017).

Notice that $\boldsymbol{\beta}$ estimates are obtained through $nlm()$ procedure with tolerance 10^{-6} in the software R, used for the whole Monte Carlo study. In Figure 1 iterations needed to compute the corresponding minimum quasi weighted DPD estimators were also obtained, with not a significant difference among the different distributions.

7.3 Comparison of minimum quasi weighted DPD estimators with pseudo minimum phi-divergence estimators

In the previous Section we have illustrated how the quasi minimum quasi weighted DPD estimators present a much more robust behavior than the maximum quasi weighted likelihood estimator for the estimation of $\boldsymbol{\beta}$ when a moderate/large sample size was considered. In Castilla et al. (2018), another family of estimators, the pseudo minimum phi-divergence estimators, was defined in the PLR model with complex survey. In particular, pseudo minimum Cressie-Read divergence estimators for positive tuning parameter λ were studied. Results of the simulation study suggested these estimators as an

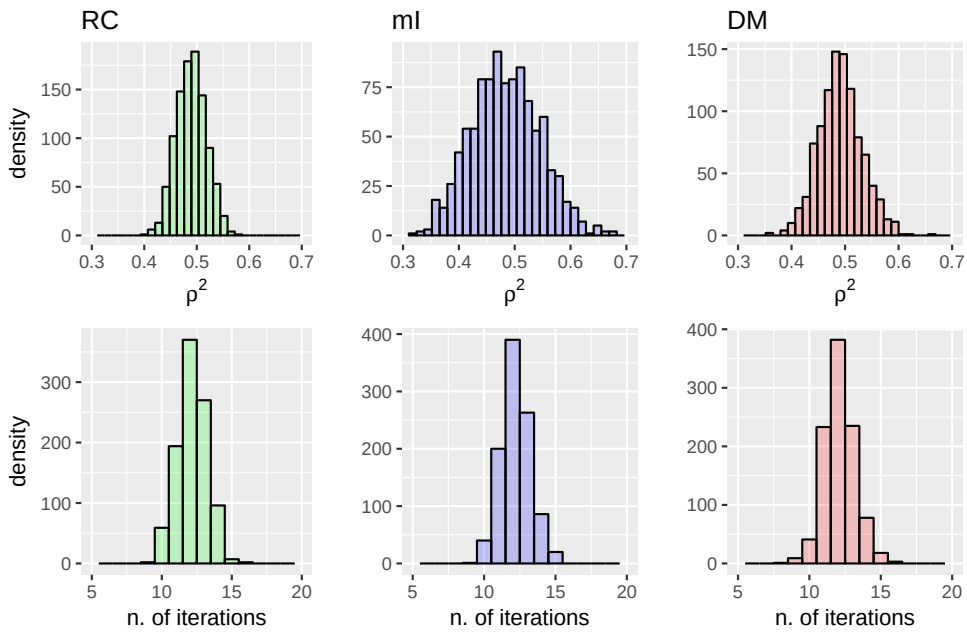

Figure 1: Density plots with estimates obtained from observations of three distributions, DM, mI, RC, when $\rho^2 = 0.5$ and $\lambda = 0.4$. $1,000$ samples.

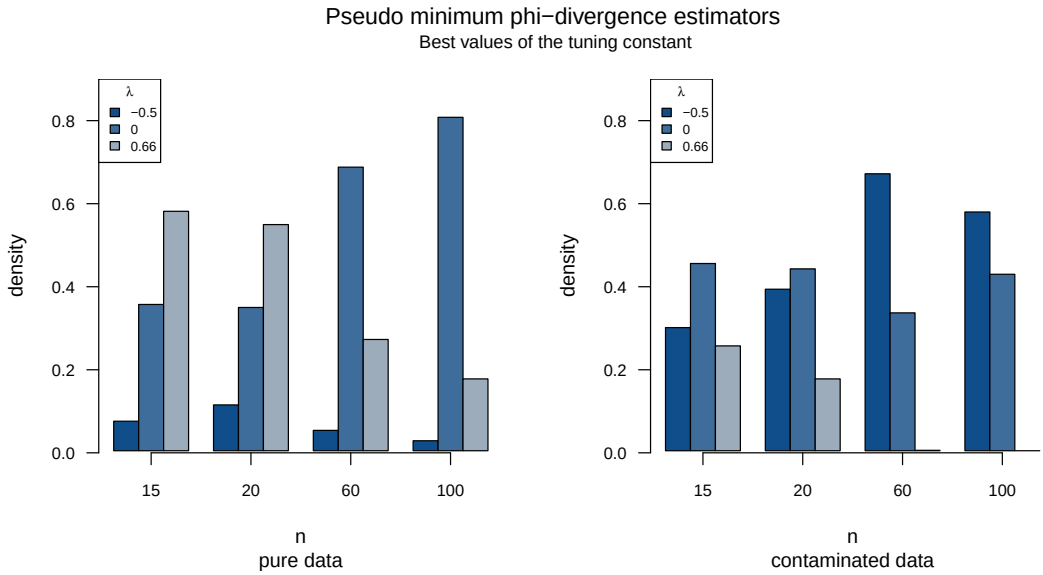

Figure 2: Pseudo minimum phi-divergence estimators: best choice of the tuning parameter, $m = 21$ and $\rho^2 = 0.5$.

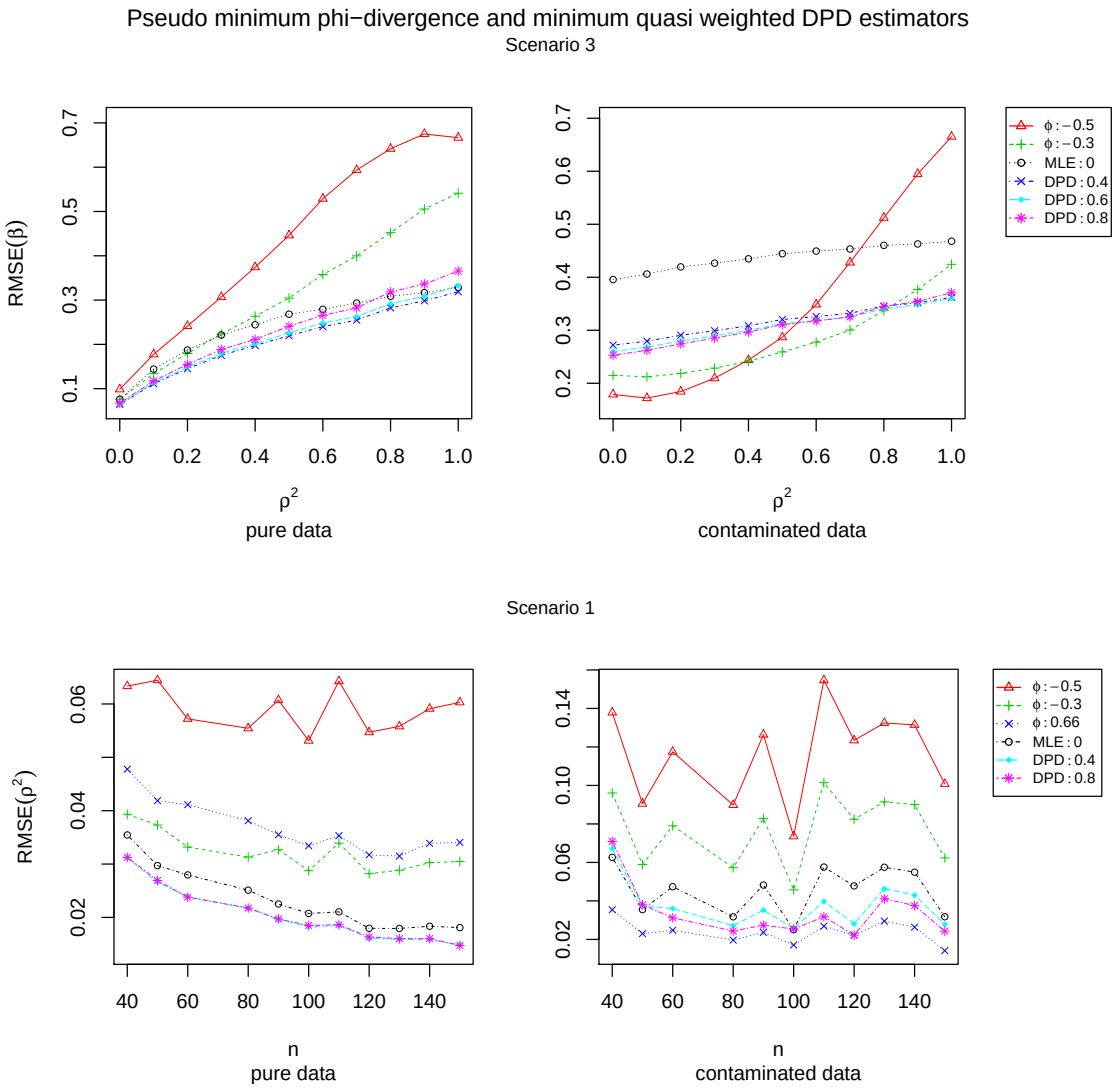

Figure 3: Comparison among pseudo minimum phidivergence and minimum quasi weighted DPD estimators

interesting alternative to maximum quasi weighted likelihood estimator in terms of efficiency when a small sample size and a large intra cluster correlation was considered. Nevertheless, the robustness of these estimators was not studied.

In this section we want to make a general empirical comparison between minimum quasi weighted DPD estimator and pseudo minimum phi-divergence estimators (with positive and negative tuning parameters in the Cressie-Read subfamily) of $\boldsymbol{\beta}$, in terms of efficiency and robustness. Although they were not studied in the cited paper of Castilla et al. (2018), phi-divergences with negative tuning

parameter and, in particular, the Hellinger distance ($\lambda = -0.5$), are known to have good robustness properties in some models. See for example Beran (1977) Lindsay (1994) and Toma (2007). This seems to happen in the context of PLR model with complex survey too, when a moderate/large sample size is considered. Nevertheless, pseudo minimum phi-divergence estimators with positive tuning parameter present an important lack of robustness. This behaviour can be summarized in Figure 2.

Let us now consider Scenario 3 of the previous section. A comparison between quasi minimum quasi weighted DPD estimators with $\lambda \in \{0.4, 0.8\}$, classical maximum quasi weighted likelihood estimator, and pseudo minimum phi-divergence estimators with $\lambda \in \{-0.5, -0.3, 0.66\}$ is made. Hellinger distance is shown to be, by far, the best choice when a contaminated scheme with low intra-cluster correlation is considered. With medium/high intra-correlation the behaviour turns to be the opposite. Minimum quasi weighted DPD estimators present a more stable behaviour, both in pure and contaminated schemes, with independence to the correlation parameter.

The same divergences are considered now in Scenario 1 for the estimation of the parameter ρ^2. The estimation of ρ^2 with pseudo minimum phi-divergence estimators is made by the method of Binder (Castilla et al., 2018) while the estimation of ρ^2 with minimum quasi weighted DPD estimators is made by the method of the estimating equations. Pseudo minimum phi-divergence estimators with negative tuning parameter are not competitive neither in the pure nor in the contaminated schemes. Pseudo minimum phi-divergence estimator with tuning parameter $\lambda = 2/3$ is a good alternative to the minimum quasi weighted DPD estimators in contaminated data. This estimator showed also a good behaviour in the paper of Castilla et al. (2018).

7.4 Performance of the Wald-type tests

With the same model as in Section 7.1 we now empirically study the robustness of the minimum quasi weighted DPD estimator based Wald-type tests for the PLR model. As in Castilla et al. (2018) we first study the level under the true null hypothesis $H_0 : \beta_{02} = 0.6$. For studying the power robustness, the true data generating parameter value is considered as $\beta_{02} = 1.08$.

Under Scenario 1, and both for pure and contaminated data, we compute observed levels and powers (measured as the proportion of test statistics exceeding the corresponding chi-square critical value), as can be seen in Figure 4. Under pure data, all levels present a very similar behaviour, while power attains their best value for classical maximum quasi weighted likelihood estimator. Under contamination, minimum quasi weighted DPD estimator based Wald-type tests for $\lambda > 0$ present a more robust behaviour than maximum quasi weighted likelihood estimator, in concordance with previous results.

8 Concluding Remarks

We have presented minimum quasi weighted DPD estimators, as a robust alternative to classical approaches, in the modeling of categorical responses with associated covariates through polytomous logistic regression models. A Wald-type family of tests based on these estimators is also presented, for the problem of testing linear hypotheses on regression coefficients. The robustness of both estimators and tests are theoretically justified in terms of the influence function, which is shown to be bounded

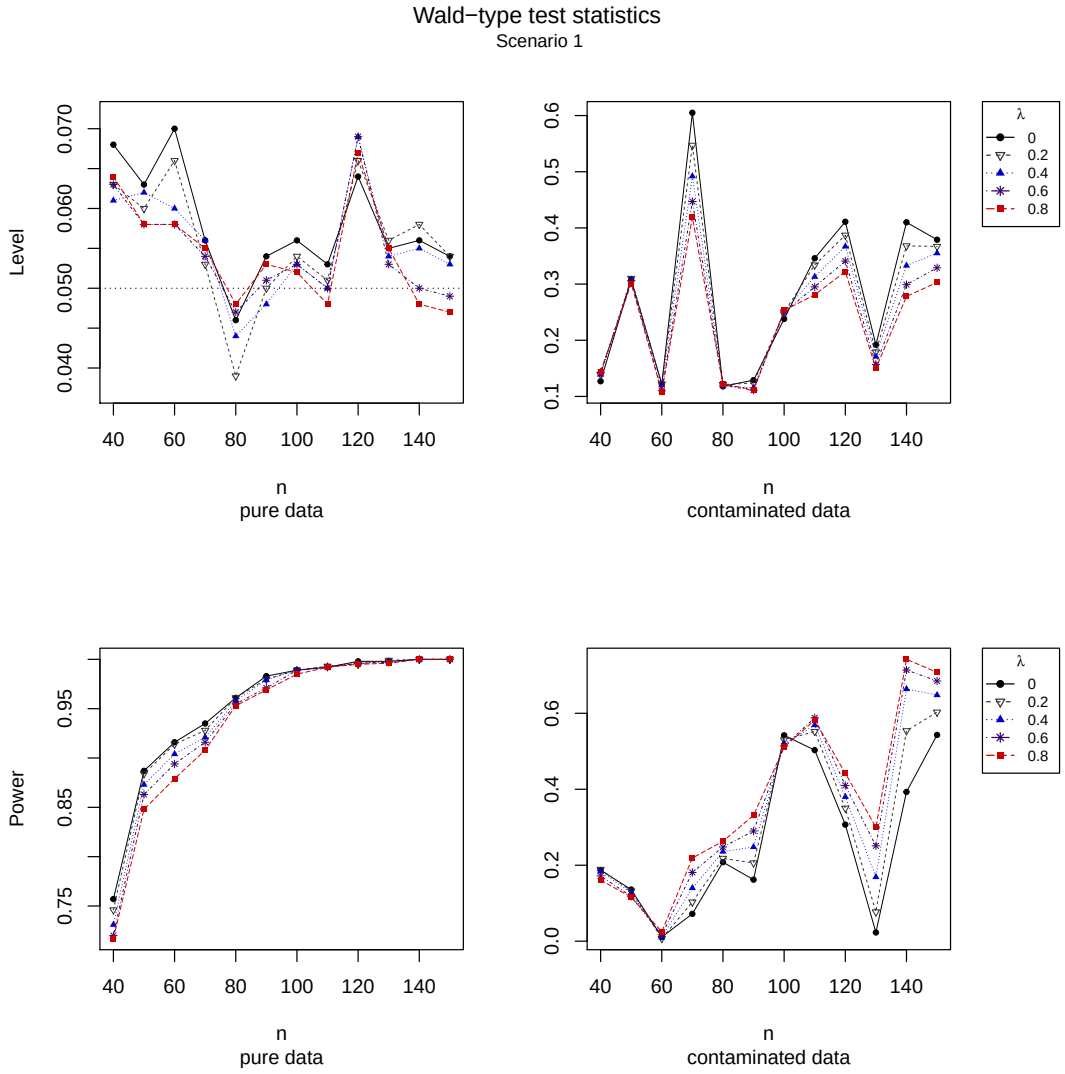

Figure 4: Estimated Levels (top) and Powers (bottom) under pure (left) and contaminated (right) data. Scenario 1.

for the new procedures. An extensive simulation study is also provided, to empirically illustrate their robustness along with a comparison with the pseudo minimum phi-divergence estimators of Castilla et al. (2018). The results clearly show how the proposed minimum quasi weighted DPD estimators seem to be the best choice for dealing with the robustness issue for a moderate sample size.

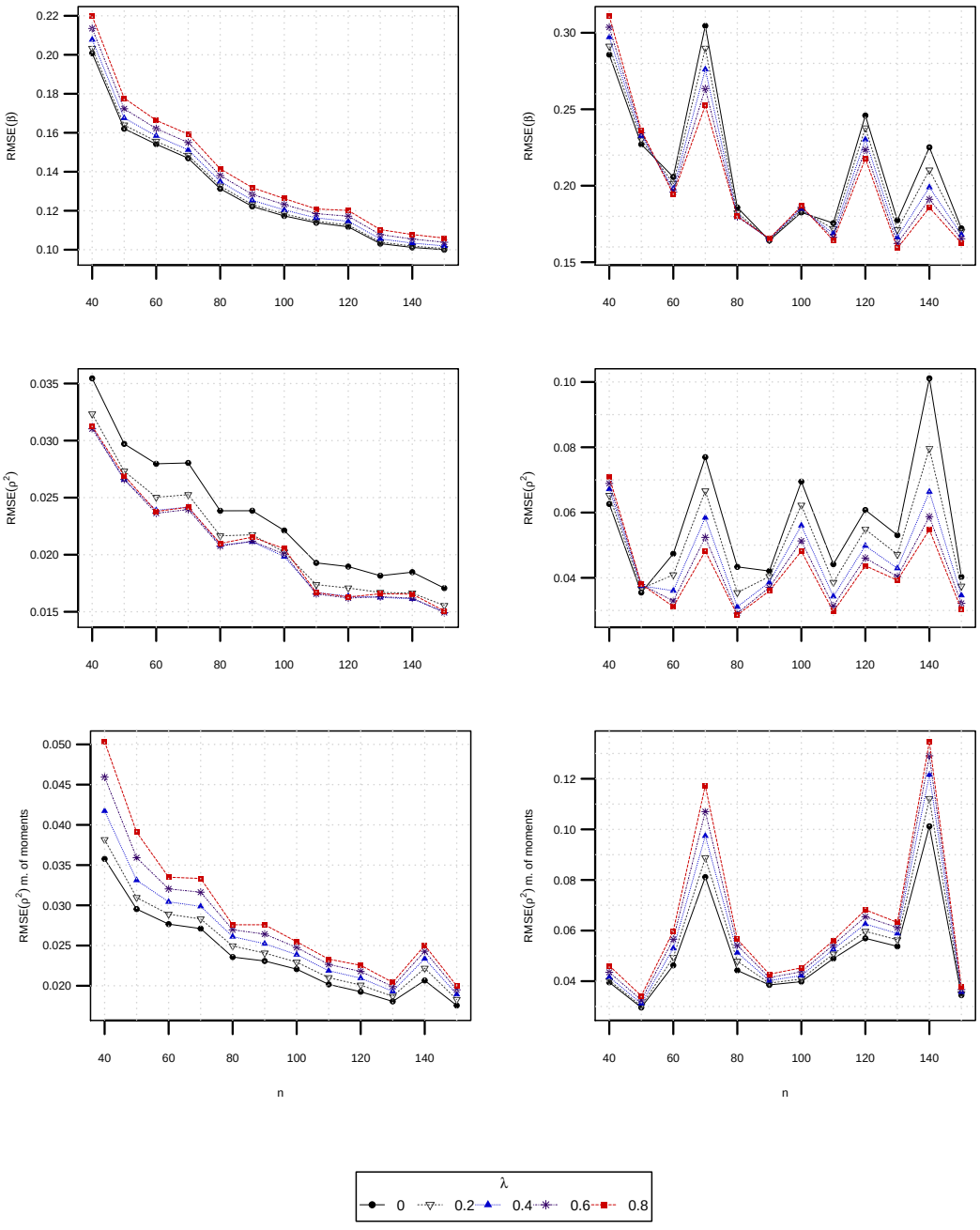

Figure 5: Scenario 1: RMSEs of minimum quasi weighted DPD estimators of $\boldsymbol{\beta}$ and ρ^2 by the equations method and the method of moments. Pure data (left) and contaminated data (right). RC distribution, $\rho^2 = 0.25$.

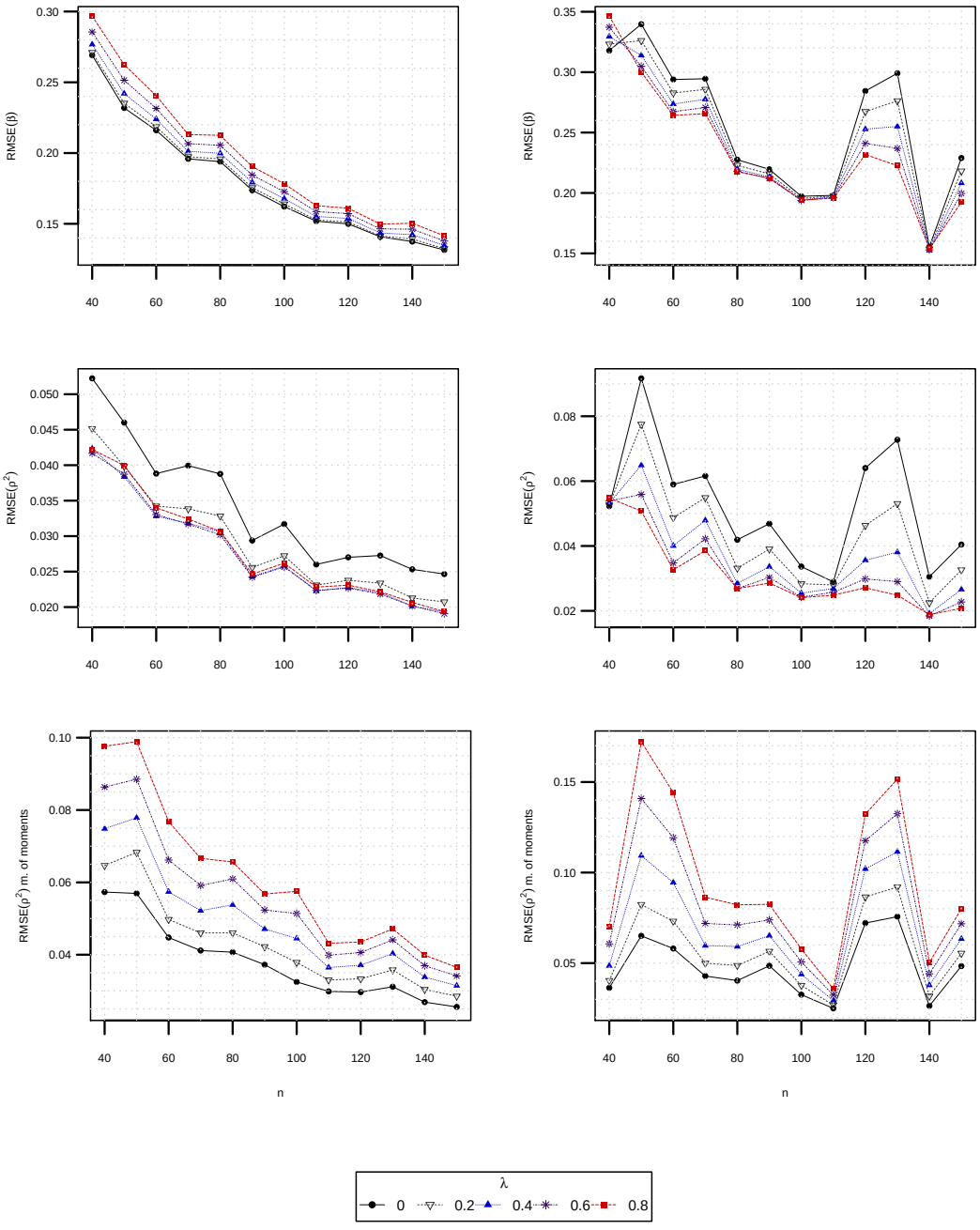

Figure 6: Scenario 1b: RMSEs of minimum quasi weighted DPD estimators of $\boldsymbol{\beta}$ and ρ^2 by the equations method and the method of moments. Pure data (left) and contaminated data (right). RC distribution, $\rho^2 = 0.50$.

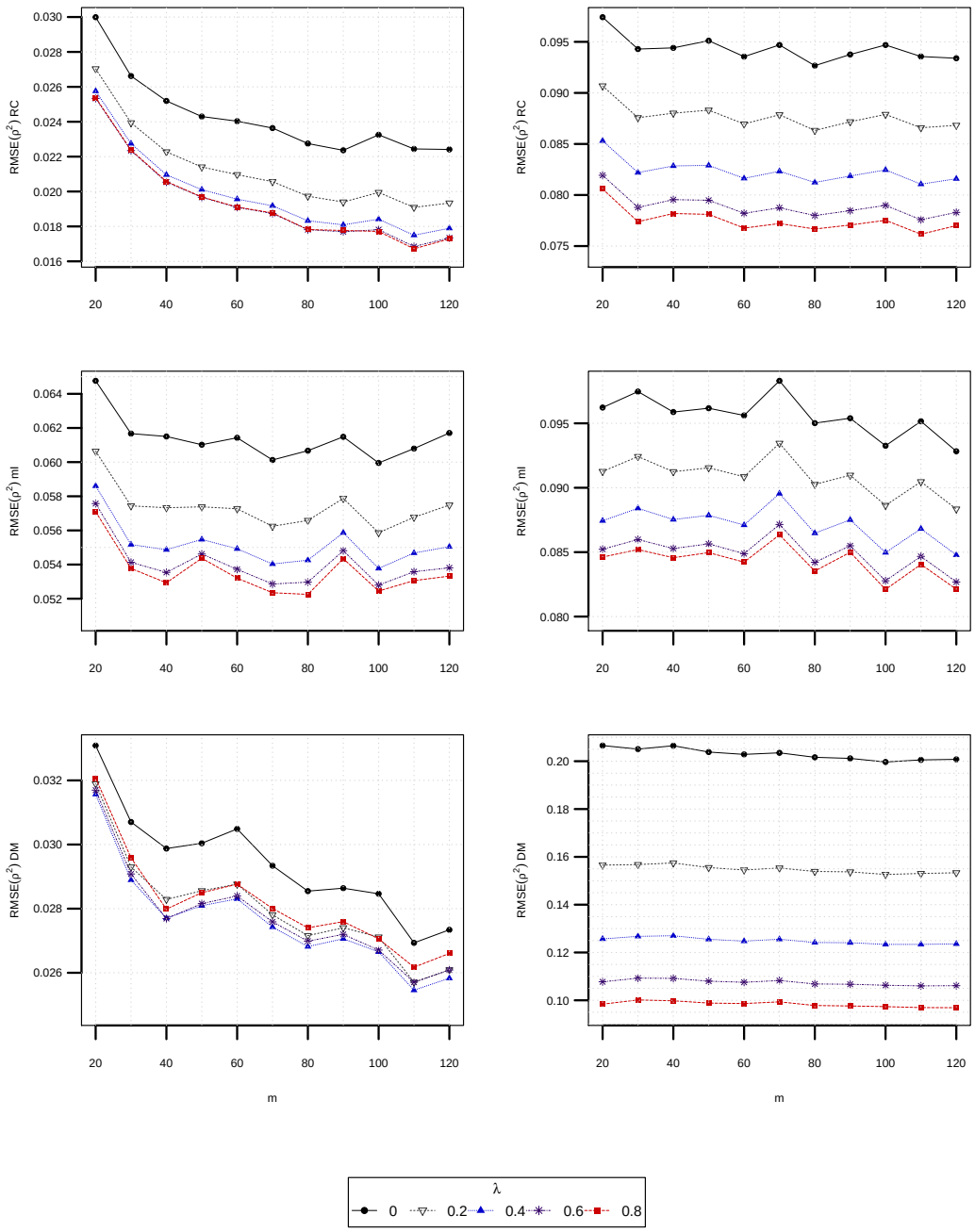

Figure 7: Scenario 2: RMSEs of ρ^2 by the equations method. Pure data (left) and contaminated data (right). RC (top), mI (middle) and DM (bottom) distribution, $\rho^2 = 0.25, n = 60$.

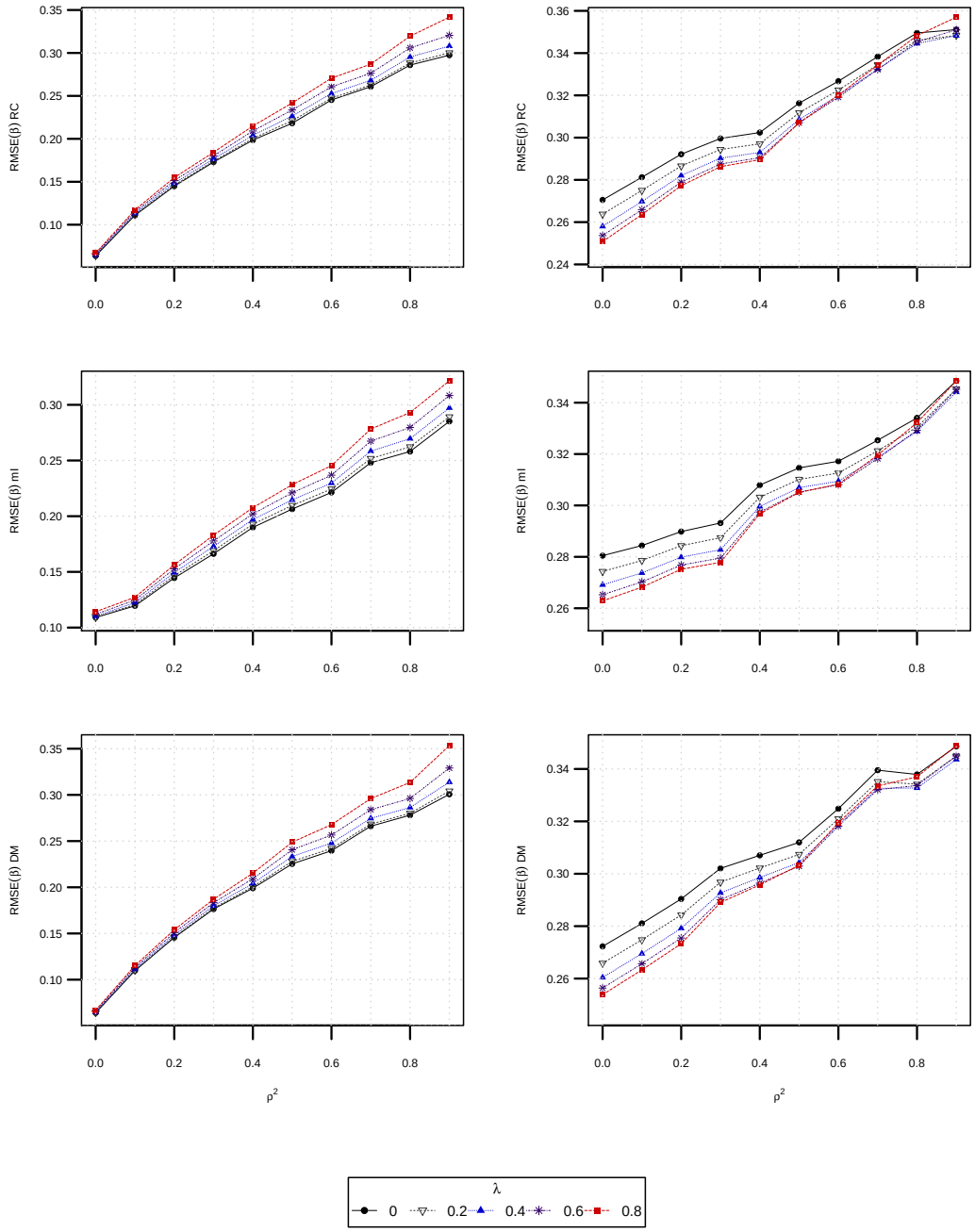

Figure 8: Scenario 3: RMSEs of minimum quasi weighted DPD estimators of $\boldsymbol{\beta}$. Pure data (left) and contaminated data (right). RC (top), mI (middle) and DM (bottom) distribution, $n = 60, m = 21$.

A Appendix

Derivation of (10) from (9)

Proof. From (9) it holds $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{u}_{hi}(\boldsymbol{\beta})=\mathbf{0}_{d(k+1)}$, with

$$\boldsymbol{u}_{hi}(\boldsymbol{\beta})=w_{hi}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{-1}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\left[\widehat{\boldsymbol{y}}_{hi}-m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\right]\otimes\boldsymbol{x}_{hi}.$$

The first term of $\boldsymbol{u}_{hi}(\boldsymbol{\beta})$ is$w_{hi}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{\widehat{\boldsymbol{y}}_{hi}-m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}$, where

$$\begin{aligned}w_{hi}\boldsymbol{\Delta}^{*}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}&=(\boldsymbol{I}_d,\mathbf{0}_d)\left[\mathrm{diag}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))-\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\,\boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})\right]\mathrm{diag}^{-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\\&=(\boldsymbol{I}_d,\mathbf{0}_d)\,(\boldsymbol{I}_{d+1}-\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\,\mathbf{1}^{T}),\end{aligned}$$

but $\mathbf{1}^{T}\{\widehat{\boldsymbol{y}}_{hi}-m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}=0$ and hence

$$\boldsymbol{u}_{hi}(\boldsymbol{\beta})=w_{hi}\left[\widehat{\boldsymbol{y}}_{hi}-m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\right]\otimes\boldsymbol{x}_{hi}.$$

■

Proof of Theorem 2

Proof. The minimum quasi weighted DPD estimator of $\boldsymbol{\beta}$, is defined as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\lambda,Q}&=\underset{\boldsymbol{\beta}\in\mathbb{R}^{d(k+1)}}{\arg\min}\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}\left\{\left(m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})-\frac{1+\lambda}{\lambda}\widehat{\boldsymbol{y}}_{hi}\right)^{T}\boldsymbol{\pi}_{hi}^{\lambda}(\boldsymbol{\beta})\right\}\\&=\underset{\boldsymbol{\beta}\in\mathbb{R}^{d(k+1)}}{\arg\min}\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}\left[\sum_{l=1}^{d+1}\left\{\left(\pi_{hil}(\boldsymbol{\beta})-\frac{1+\lambda}{\lambda}\frac{\widehat{y}_{hil}}{m_{hi}}\right)\pi_{hil}^{\lambda}(\boldsymbol{\beta})\right\}\right]\end{aligned}$$

which can also be obtained by solving the system of equations $\boldsymbol{u}_{\lambda}(\boldsymbol{\beta})=\mathbf{0}_{d(k+1)}$ where

$$\begin{aligned}-\frac{1}{\lambda+1}\frac{\partial}{\partial\boldsymbol{\beta}}\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}\left\{\left(m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})-\frac{1+\lambda}{\lambda}\widehat{\boldsymbol{y}}_{hi}\right)^{T}\boldsymbol{\pi}_{hi}^{\lambda}(\boldsymbol{\beta})\right\}&=\mathbf{0}_{d(k+1)},\\\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}m_{hi}\sum_{l=1}^{d+1}\left\{\frac{\widehat{y}_{hil}}{m_{hi}}-\pi_{hil}(\boldsymbol{\beta})\right\}\pi_{hil}^{\lambda-1}(\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}}\pi_{hil}(\boldsymbol{\beta})&=\mathbf{0}_{d(k+1)},\\\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}m_{hi}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\pi}_{hi}^{T}(\boldsymbol{\beta})\,\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{\frac{\widehat{\boldsymbol{y}}_{hi}}{m_{hi}}-\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right]&=\mathbf{0}_{d(k+1)}.\end{aligned}$$

Now, taking into account that

$$\frac{\partial\pi_{hil}(\boldsymbol{\beta})}{\partial\beta_{uv}}=x_{iu}\pi_{hil}(\boldsymbol{\beta})\left\{\delta_{jv}-\pi_{hiv}(\boldsymbol{\beta})\right\},\qquad u=1,...,k,v=1,...,d,$$

we get

$$\frac{\partial \boldsymbol{\pi}_{hi}^T(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = (\boldsymbol{I}_d, \mathbf{0}_d)\, \boldsymbol{\Delta}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})) \otimes \boldsymbol{x}_i = \boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})) \otimes \boldsymbol{x}_i, \tag{37}$$

and hence the system of equations becomes

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left[w_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{\widehat{\boldsymbol{y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right] \otimes \boldsymbol{x}_{hi} = \mathbf{0}_{d(k+1)}.$$

■

Proof of Theorem 3

Proof. By following formulas (3.3) and (3.4) of Ghosh and Basu (2013), it holds

$$\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}) = \mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})\right], \quad \boldsymbol{\Psi}_\lambda(\boldsymbol{\beta}) = \mathrm{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{U}_\lambda^T(\boldsymbol{\beta}, \boldsymbol{X})\right],$$

which can be expressed as a limit of

$$\boldsymbol{\Omega}_{n,\lambda}(\boldsymbol{\beta}) = \widehat{\mathrm{E}}[\mathrm{Var}[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X}]], \quad \boldsymbol{\Psi}_{n,\lambda}(\boldsymbol{\beta}) = \widehat{\mathrm{E}}\left[\mathrm{E}\left[-\tfrac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{U}_\lambda^T(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X}\right]\right],$$

where the summands given in (14) and (15) are

$$\boldsymbol{\Omega}_{hi,\lambda}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{hi}) = \mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{x}_{hi})\right], \quad \boldsymbol{\Psi}_{hi,\lambda}(\boldsymbol{\beta}) = \mathrm{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{U}_\lambda^T(\boldsymbol{\beta}, \boldsymbol{x}_{hi})\right].$$

With respect to $\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}) = \mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})\right] = \mathrm{E}\left[\mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X}\right]\right] + \mathrm{Var}\left[\mathrm{E}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X}\right]\right]$, the first summand estimated from the sample is

$$\widehat{\mathrm{E}}[\mathrm{Var}[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X}]] = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{x}_{hi})\right]\widehat{\Pr}(\boldsymbol{X} = \boldsymbol{x}_{hi}) = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{x}_{hi})\right],$$

and the second one

$$\widehat{\mathrm{Var}}\left[\mathrm{E}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X}\right]\right] = \mathrm{Var}\left[\sum_{h=1}^{H}\sum_{i=1}^{n_h}E\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{x}_{hi})\right]\widehat{\Pr}(\boldsymbol{X} = \boldsymbol{x}_{hi})\right] = \mathbf{0}_{d(k+1)},$$

for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, since $\mathrm{E}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}_0, \boldsymbol{x}_{hi})\right] = \mathbf{0}_{d(k+1)}$ from $\mathrm{E}[\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}_0)] = \mathbf{0}_{d(k+1)}$. On the other hand,

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{x}_{hi})\right] &= \mathrm{Var}\left[\left[w_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\{\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right] \otimes \boldsymbol{x}_{hi}\right] \\ &= \left(w_{hi}^2\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\mathrm{Var}[\widehat{\boldsymbol{Y}}_{hi}]\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\right) \otimes \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T.\end{aligned}$$

With respect to $\boldsymbol{\Psi}_\lambda(\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{U}_\lambda^T(\boldsymbol{\beta},\boldsymbol{X})\right]$, estimated from the sample is given by

$$
\begin{aligned}
&\widehat{\mathrm{E}}\left[\mathrm{E}\left[-\tfrac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{U}_\lambda^T(\boldsymbol{\beta},\boldsymbol{X})|\boldsymbol{X}\right]\right] \\
&= -\sum_{h=1}^{H}\sum_{i=1}^{n_h}\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{U}_\lambda(\boldsymbol{\beta},\boldsymbol{x}_{hi})\right]\widehat{\Pr}(\boldsymbol{X}=\boldsymbol{x}_{hi}) \\
&= -\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left[w_{hi}\frac{\partial}{\partial\boldsymbol{\beta}}\left[\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\right]\mathrm{E}[\widehat{\boldsymbol{Y}}_{hi}-m_{hi}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})]\right]^T\otimes\boldsymbol{x}_{hi}^T \\
&+\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left[w_{hi}m_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\right]^T\otimes\boldsymbol{x}_{hi}^T \\
&= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left[w_{hi}m_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\right]^T\otimes\boldsymbol{x}_{hi}^T \\
&= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left[w_{hi}m_{hi}\boldsymbol{\Delta}^*(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\mathrm{diag}^{\lambda-1}\{\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\}\boldsymbol{\Delta}^{*T}(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta}))\right]\otimes\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T,
\end{aligned}
$$

where the third equality is again true for $\boldsymbol{\beta}=\boldsymbol{\beta}_0$. The derivations are omitted for $\lambda=0$, in which case similar ideas as $\lambda>0$ might be followed. ■

Proof of Theorem 14

Proof. We have

$$
\begin{aligned}
Po_{W_n(\widehat{\boldsymbol{\beta}}_{\lambda,Q})}(\boldsymbol{\beta}^*) &= \Pr\left(W_n\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}\right)>\chi^2_{r,\alpha}\right)=\Pr\left(n\left(l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})-l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)>\chi^2_{r,\alpha}-nl_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right) \\
&= \Pr\left(\sqrt{n}\left(l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})-l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)>\frac{\chi^2_{r,\alpha}}{\sqrt{n}}-\sqrt{n}l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right).
\end{aligned}
$$

Now we are going to get the asymptotic distribution of the random variable $\sqrt{n}\left(l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})-l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)$.

It is clear that $l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$ and $l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\boldsymbol{\beta}^*)$ have the same asymptotic distribution because $\widehat{\boldsymbol{\beta}}_{\lambda,Q}\xrightarrow[n\to\infty]{P}\boldsymbol{\beta}^*$. A first Taylor expansion of $l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\boldsymbol{\beta}^*)$ at $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$ around $\boldsymbol{\beta}^*$ gives

$$
l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\boldsymbol{\beta}^*)-l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)=\left.\frac{\partial l_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{\beta}^*)+o_p\left(\left\|\widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{\beta}^*\right\|\right).
$$

Therefore,

$$
\sqrt{n}\left(l_{\widehat{\boldsymbol{\beta}}_{\lambda,Q}}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})-l_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)\xrightarrow[n\to\infty]{L}N\left(0,\sigma^2(\boldsymbol{\beta}^*)\right)
$$

being

$$\sigma^2\left(\boldsymbol{\beta}^*\right) = \left.\frac{\partial l_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \widehat{\mathbf{Q}}_{n,\lambda}\left(\boldsymbol{\beta}^*\right) \left.\frac{\partial l_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

■

Proof of Corollary 4

Proof. It is based on the weak consistency of $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$ as well as in the continuity with respect to $\boldsymbol{\beta}$ of the elements of different matrices. Based on Theorem 3, it holds $\lim_{n\to\infty} \mathrm{E}[\widehat{\boldsymbol{\beta}}_{\lambda,Q}] = \mathbf{0}_{d(k+1)}$ and $\lim_{n\to\infty} \mathrm{Var}[\widehat{\boldsymbol{\beta}}_{\lambda,Q}] = \boldsymbol{O}_{d(k+1)\times d(k+1)}$ and hence $\widehat{\boldsymbol{\beta}}_{\lambda,Q} \xrightarrow[n\to\infty]{P} \boldsymbol{\beta}_0$.
■

Proof of Theorem 6

Proof. Let

$$\begin{aligned}\widehat{\boldsymbol{\Omega}}_{n,\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) &= \widehat{\mathrm{E}}[\mathrm{Var}[\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})|\boldsymbol{X}]] \\ &= \widehat{\mathrm{E}}[\mathrm{E}[\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})\boldsymbol{U}_\lambda^T(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})|\boldsymbol{X}] - \mathrm{E}[\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})|\boldsymbol{X}]\mathrm{E}[\boldsymbol{U}_\lambda^T(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})|\boldsymbol{X}]]\end{aligned}$$

be another possible consistent estimator of $\boldsymbol{\Omega}_\lambda\left(\boldsymbol{\beta}\right)$, alternative to $\boldsymbol{\Omega}_{n,\lambda}\left(\boldsymbol{\beta}\right)$. Since

$$\widehat{\mathrm{E}}\left[\mathrm{E}[\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})|\boldsymbol{X}]\mathrm{E}[\boldsymbol{U}_\lambda^T(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})|\boldsymbol{X}]]\right] = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \mathrm{E}[\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{x}_{hi})]\mathrm{E}[\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{x}_{hi})]\widehat{\Pr}(\boldsymbol{X} = \boldsymbol{x}_{hi}),$$

and $\mathrm{E}[\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda,Q}, \boldsymbol{X})|\boldsymbol{x}_{hi}] = 0$, for being $\mathrm{E}[\widehat{\boldsymbol{Y}}_{hi} - m_{hi}\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})] = \mathbf{0}_{k+1}$, for all $(h,i) \in \{1,...,H\} \times \{1,...,m_{hi}\}$ according to the PLR model with complex design, it holds

$$\widehat{\boldsymbol{\Omega}}_{n,\lambda}(\boldsymbol{\beta}) = \widehat{\mathrm{E}}[\mathrm{E}[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X})\boldsymbol{U}_\lambda^T(\boldsymbol{\beta}, \boldsymbol{X})|\boldsymbol{X}] = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} \boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{x}_{hi})\boldsymbol{U}_\lambda^T(\boldsymbol{\beta}, \boldsymbol{x}_{hi})$$

is another possible consistent estimator of $\boldsymbol{\Omega}_\lambda\left(\boldsymbol{\beta}\right)$, alternative to $\boldsymbol{\Omega}_{n,\lambda}\left(\boldsymbol{\beta}\right)$. ■

Proof of Corollary 7

Proof. Section a) is straightforward considering the expressions of matrix $\mathbf{G}_{n,\lambda}(\boldsymbol{\beta})$ and the related ones. For section b) we consider the vector

$$\boldsymbol{Z}_{hi}^*\left(\boldsymbol{\beta}\right) = \sqrt{m}\Delta^{-1/2}\left(\boldsymbol{\pi}_{hi}^*\left(\boldsymbol{\beta}\right)\right)\left(\tfrac{\widehat{\boldsymbol{Y}}_{hi}^*}{m} - \boldsymbol{\pi}_{hi}^*\left(\beta\right)\right).$$

Taking into account that $\mathrm{Var}[\widehat{\boldsymbol{Y}}_{hi}^*] = m\nu\boldsymbol{\Delta}\left(\boldsymbol{\pi}_{hi}^*\left(\beta\right)\right)$, we have

$$\mathrm{E}\left[\boldsymbol{Z}_{hi}^*\left(\boldsymbol{\beta}\right)\right] = \mathbf{0} \text{ and } \mathrm{Var}\left[\boldsymbol{Z}_{hi}^*\left(\boldsymbol{\beta}\right)\right] = \nu\boldsymbol{I}_d,$$

for $h = 1, ..., H$ and $i = 1, ..., n_h$. An unbiased estimator of $\mathrm{Var}\left[\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\right]$ is

$$\widehat{\mathrm{Var}}\left[\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\right] = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{Z}^*_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\boldsymbol{Z}^{*T}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}),$$

i.e. $\mathrm{E}\left[\widehat{\mathrm{Var}}\left[\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\right]\right] = \mathrm{Var}\left[\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\right]$ for which the trace is

$$\begin{aligned}
\mathrm{E}\left[\mathrm{trace}\left(\widehat{\mathrm{Var}}\left[\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\right]\right)\right] &= \mathrm{trace}\left(\mathrm{Var}\left[\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\right]\right), \\
\mathrm{E}\left[\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\boldsymbol{Z}^{*T}_{hi}(\boldsymbol{\beta})\right] &= \nu d, \\
\mathrm{E}\left[\frac{1}{nd}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{Z}^*_{hi}(\boldsymbol{\beta})\boldsymbol{Z}^{*T}_{hi}(\boldsymbol{\beta})\right] &= \nu.
\end{aligned}$$

Since $\widehat{\boldsymbol{\beta}}_{\lambda,Q}$ is consistent of $\boldsymbol{\beta}$, it holds that $\frac{1}{nd}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{Z}^*_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\boldsymbol{Z}^{*T}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})$ is consistent of ν, but

$$\frac{1}{nd}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\boldsymbol{Z}^*_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})\boldsymbol{Z}^{*T}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}) = \frac{1}{nd}\sum_{h=1}^{H}\sum_{i=1}^{n_h}(\widehat{\boldsymbol{Y}}^*_{hi} - m\boldsymbol{\pi}^*_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))^T\frac{1}{m}\Delta^{-1}\left(\boldsymbol{\pi}^*_{hi}(\boldsymbol{\beta})\right)(\widehat{\boldsymbol{Y}}^*_{hi} - m\boldsymbol{\pi}^*_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})),$$

is in principle different in shape in comparison with (20) with

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{d+1}\frac{(\widehat{Y}_{hij} - m\pi_{hij}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))^2}{m\pi_{hij}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}(\widehat{\boldsymbol{Y}}_{hi} - m\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))^T\frac{1}{m}\Delta^{-}\left(\boldsymbol{\pi}^*_{hi}(\boldsymbol{\beta})\right)(\widehat{\boldsymbol{Y}}_{hi} - m\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})),$$

where $\Delta^{-}\left(\boldsymbol{\pi}^*_{hi}(\boldsymbol{\beta})\right) = \mathrm{diag}^{-1}\left(\boldsymbol{\pi}_{hi}(\boldsymbol{\beta})\right)$. To assure that both expressions are equivalent we count with the result related to invariance of quadratic forms with generalized variances, Lemma 1a, of Moore (1977), from which is concluded that

$$\begin{aligned}
&(\widehat{\boldsymbol{Y}}_{hi} - m\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))^T\frac{1}{m}\Delta^{-}\left(\boldsymbol{\pi}^*_{hi}(\boldsymbol{\beta})\right)(\widehat{\boldsymbol{Y}}_{hi} - m\boldsymbol{\pi}_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})) \\
&= (\widehat{\boldsymbol{Y}}^*_{hi} - m\boldsymbol{\pi}^*_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q}))^T\frac{1}{m}\Delta^{-1}\left(\boldsymbol{\pi}^*_{hi}(\boldsymbol{\beta})\right)(\widehat{\boldsymbol{Y}}^*_{hi} - m\boldsymbol{\pi}^*_{hi}(\widehat{\boldsymbol{\beta}}_{\lambda,Q})).
\end{aligned}$$

■

Proof of Theorem 8

Proof. The details are omitted for being very similar to Theorem 3 conditioning within each stratum, i.e.

$$\begin{aligned}
\boldsymbol{\Omega}^{(h)}_{i,\lambda}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{hi}) &= \widehat{\mathrm{E}}[\mathrm{Var}[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}_h)|\boldsymbol{X}_h]], \quad \boldsymbol{\Omega}^{(h)}_{\lambda}(\boldsymbol{\beta}) = \mathrm{Var}\left[\boldsymbol{U}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}_h)\right], \\
\boldsymbol{\Psi}^{(h)}_{i,\lambda}(\boldsymbol{\beta}) &= -\widehat{\mathrm{E}}\left[\mathrm{E}\left[\tfrac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{U}^T_\lambda(\boldsymbol{\beta}, \boldsymbol{X}_h)|\boldsymbol{X}_h\right]\right], \quad \boldsymbol{\Psi}^{(h)}_{\lambda}(\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{U}^T_\lambda(\boldsymbol{\beta}, \boldsymbol{X}_h)\right].
\end{aligned}$$

■

Proof of Theorem 16

Proof. We have

$$\begin{aligned}\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{l} &= \boldsymbol{M}^T\ \boldsymbol{\beta}_n - l + \boldsymbol{M}^T\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}-\ \boldsymbol{\beta}_n\right)\\ &= \boldsymbol{M}^T\ \boldsymbol{\beta}_0 + \boldsymbol{M^T}n^{-1/2}\boldsymbol{d} - \boldsymbol{l} + \boldsymbol{L^T}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda,Q}}-\ \boldsymbol{\beta_n}\right)\\ &= \boldsymbol{M}\ ^{\boldsymbol{T}}n^{-1/2}\boldsymbol{d} + \boldsymbol{M^T}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda,Q}}-\ \boldsymbol{\beta_n}\right).\end{aligned}$$

Therefore,

$$\boldsymbol{M}^T\ \widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{l} = \boldsymbol{M^T}n^{-1/2}\boldsymbol{d} + \boldsymbol{M^T}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda,Q}}-\ \boldsymbol{\beta_n}\right).$$

We know, under $H_{1,n}$ that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{\beta_n}\right)\ \underset{n\longrightarrow\infty}{\overset{\mathcal{L}}{\longrightarrow}}\ \mathcal{N}\left(\mathbf{0},\ \mathbf{Q}_\lambda(\boldsymbol{\beta_0})\right)$$

and

$$\sqrt{n}\left(\boldsymbol{M}^T\ \widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{c}\right)\ \underset{n\longrightarrow\infty}{\overset{\mathcal{L}}{\longrightarrow}}\ \mathcal{N}\left(M^{\boldsymbol{T}}\boldsymbol{d},\ M^{\boldsymbol{T}}\boldsymbol{\Psi_\lambda}\left(\boldsymbol{\beta_0}\right)\boldsymbol{M}\right).$$

The Wald-type test statistics can be written as the quadratic form $W_n\left(\widehat{\boldsymbol{\beta}}_{\lambda,Q}\right)=\boldsymbol{Z}^T\boldsymbol{Z}$ with

$$\boldsymbol{Z}=\sqrt{n}\left[M^{\boldsymbol{T}}\boldsymbol{\Psi_\lambda}\left(\boldsymbol{\beta_0}\right)\boldsymbol{M}\right]^{-1/2}\left(\boldsymbol{M}^T\ \widehat{\boldsymbol{\beta}}_{\lambda,Q}-\boldsymbol{l}\right)$$

and

$$\boldsymbol{Z}\ \underset{n\longrightarrow\infty}{\overset{\mathcal{L}}{\longrightarrow}}\ \mathcal{N}\left(\left[M^{\boldsymbol{T}}\boldsymbol{\Psi_\lambda}\left(\boldsymbol{\beta_0}\right)\boldsymbol{M}\right]^{-1/2}\boldsymbol{M^T}\boldsymbol{d},\boldsymbol{I}_{r\times r}\right),$$

where $\boldsymbol{I}$ is the identity $r\times r$ matrix. The application of the previous gives i). The noncentrality parameter is $\boldsymbol{d}^T\boldsymbol{M}\left[M^{\boldsymbol{T}}\boldsymbol{\Psi_\lambda}\left(\boldsymbol{\beta_0}\right)\boldsymbol{M}\right]^{-1}\boldsymbol{M^T}\boldsymbol{d}$**.** Result ii) follows using relation (27). ■

References

[1] Agresti, A. (2002). *Categorical Data Analysis* (Second Edition). John Wiley & Sons.

[2] Alonso-Revenga, J. M., Martín, N. and Pardo, L. (2017). New improved estimators for overdispersion in models with clustered multinomial data and unequal cluster sizes, *Statistics and Computing* **27**, 193-217.

[3] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, **85**, 549–559

[4] Basu, A. Ghosh, A. Mandal, N. Martin and L. Pardo (2017). A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electon. J. Stat.*, **11**, 2741–2772

[5] Basu, A. Ghosh, A., Martin and L. Pardo (2018). Robust Wald-type tests for non-homogeneous observations based on the minimum density power divergence estimator. *Metrika*, **81(5)**, 493–522.

[6] Beran, Rudolf (1977). Minimum Hellinger Distance Estimates for Parametric Models. *Ann. Statist.* **5**, no. 3, 445–463.

[7] Binder, D. A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review, **51**, 279–292.*

[8] Castilla, E., Martin, N. and Pardo, L. (2018). Minimum phi-divergence estimators for multinomial logistic regression with complex sample design. *Advances in Statistical Analysis*, **102**, 381–411.

[9] Castilla, E., Ghosh, A., Martin, N. and Pardo, L. (2019). New robust statistical procedures for polytomous logistic regression models. *Biometrics.* **74**, 1282-1291.

[10] Engel, J. (1988). Polytomous logistic regression. *Statistica Neerlandica,* **42,** 233–252.

[11] Ghosh, A., and Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, **7**, 2420–2456.

[12] Ghosh, A. and Basu, A. (2015). Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: the density power divergence approach. *Journal of applied statitsics*, **42**, 2056–2072.

[13] Ghosh, A., and Basu, A. (2016). Robust Estimation in Generalized Linear Models : The Density Power Divergence Approach. *TEST*, **25**, 269–290.

[14] Ghosh, A., and Basu, A. (2018). Robust Bounded Influence Tests for Independent but Non-Homogeneous Observations. *Statistica Sinica*, **28**, 1133–1155.

[15] Gupta, A. K., Kasturiratna, D., Nguyen, T. and Pardo, L. (2006). A new family of BAN estimators for polytomous logistic regression models based on density power divergence measures. Statistical Methods & Applications, **15**, 159–176.

[16] Gupta, A. K.; Nguyen, T.; Pardo, L. (2008). Residuals for polytomous logistic regression models based on density power divergences test statistics. *Statistics*, **42**, 495–514.

[17] Hampel, F. R., Ronchetti, E., Rousseeuw, P. J., and Stahel W.(1986). Robust Statistics: The Approach Based on Influence Functions. *New York, USA: John Wiley & Sons.*

[18] Huber, P. J. (1983). Minimax aspects of bounded-influence regression (with discussion). *Journal of the American Statistical Association*, **69**, 383-393.

[19] Lesaffre, E.and Albert, A. (1989). Multiple-group logistic regression diagnostic. *Applied Statistics,* **38,** 425–440.

[20] Lindsay, Bruce G. (1994). Efficiency Versus Robustness: The Case for Minimum Hellinger Distance and Related Methods. *Ann. Statist.* **22**, no. 2, 1081–1114.

[21] Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: an overview and a survey of recent developments. With discussion and a rejoinder by the authors. *Test*, **14**, 1–73.

[22] McCullagh, P. (1980). Regression models for ordinary data. *Journal of the Royal Statistical Society*-Series B, **42,** 109–142.

[23] Morel, G. (1989). Logistic regression under Complex Survey Designs. *Survey Methodology, **15**, 203–223.*

[24] Morel, J. G. and Koehler, K. J. (1995). A OneStep GaussNewton Estimator for Modelling Categorical Data with Extraneous Variation. *Journal of the Royal Statistical Society: Series C*, **44(2)**, 187–200.

[25] Morel, G. and Neerchal, N. K. (2012). Overdispersion Models in SAS. SAS Institute.

[26] Pardo, L. (2005). *Statistical Inference Based on Divergence Measures. Statistics: Texbooks and Monographs.* Chapman & Hall/CRC, New York.

[27] Raim, A. M. , Neerchal, N. K. and Morel, J. G. (2015). Modeling overdispersion in R. Technical Report HPCI-2015-1 UMBCH High Performance Computing Facility, University of Maryland, Baltimore Country.

[28] Roberts, G., Rao, J.N.K. and Kumer, S. (1987). Logistic Regression Analysis of Sample Survey Data, *Biometrika*, **74**, 1–12.

[29] Toma, A. (2007). Minimum Hellinger distance estimators for some multivariate models: influence functions and breakdown point results. *C. R. Acad. Sci. Paris*, Ser. I 345, 353–358.

[30] Warwick, J. and Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, **75**, 581–588.

[31] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.