Electronic Journal of Statistics
Vol. 11 (2017) 2741–2772
ISSN: 1935-7524
DOI: 10.1214/17-EJS1295

A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator*

Ayanendranath Basu and Abhik Ghosh

Interdisciplinary Statistical Research Unit
Indian Statistical Institute
e-mail: ayanbasu@isical.ac.in; abhianik@gmail.com

Abhijit Mandal

Department of Mathematics
Wayne State University
e-mail: abhijit.mandal@wayne.edu

Nirian Martín* and Leandro Pardo*

Department of Statistics and O.R.
Complutense University of Madrid
e-mail: nirian@estad.ucm.es; lpardo@mat.ucm.es

Abstract: In this paper a robust version of the classical Wald test statistics for linear hypothesis in the logistic regression model is introduced and its properties are explored. We study the problem under the assumption of random covariates although some ideas with non random covariates are also considered. A family of robust Wald type tests are considered here, where the minimum density power divergence estimator is used instead of the maximum likelihood estimator. We obtain the asymptotic distribution and also study the robustness properties of these Wald type test statistics. The robustness of the tests is investigated theoretically through the influence function analysis as well as suitable practical examples. It is theoretically established that the level as well as the power of the Wald-type tests are stable against contamination, while the classical Wald type test breaks down in this scenario. Some classical examples are presented which numerically substantiate the theory developed. Finally a simulation study is included to provide further confirmation of the validity of the theoretical results established in the paper.

MSC 2010 subject classifications: Primary 62F35, 662F05.
Keywords and phrases: Influence function, logistic regression, minimum density power divergence estimators, random explanatory variables, robustness, Wald-type test statistics.

Received July 2016.

*This research is partially supported by Grant MTM2015-67057-P from Ministerio de Economia y Competitividad (Spain)

1. Introduction

Experimental settings often include dichotomous response data, wherein a Bernoulli model may be assumed for independent response variables $Y_1, ..., Y_n$, with

$$\Pr(Y_i = 1) = \pi_i \text{ and } \Pr(Y_i = 0) = 1 - \pi_i,\ i = 1, ..., n.$$

In many cases, a series of explanatory variables $x_{i0}, ..., x_{ik}$ may be associated with each Y_i ($x_{i0} = 1$, $x_{ij} \in \mathbb{R}$, $i = 1, ..., n$, $j = 1, ..., k$, $k < n$). We shall assume that the binomial parameter, π_i, is linked to the linear predictor $\sum_{j=0}^{k} \beta_j x_{ij}$ via the logit function, i.e.,

$$\text{logit}\,(\pi_i) = \sum_{j=0}^{k} \beta_j x_{ij}, \tag{1.1}$$

where $\text{logit}(p) = \log(p/(1-p))$. In the following, we shall denote the binomial parameter π_i, by

$$\pi_i = \pi(\boldsymbol{x}_i^T \boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}},\ i = 1, ..., n, \tag{1.2}$$

where $\boldsymbol{x}_i^T = (x_{i0}, ..., x_{ik})$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)^T$ is a $(k+1)$-dimensional vector of unknown parameters with $\beta_i \in (-\infty, \infty)$. The "design matrix", $\mathbb{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$, is assumed to be full rank ($rank(\mathbb{X}) = k + 1$), without any loss of generality.

Let $\boldsymbol{M}$ be any matrix of r rows and $k+1$ columns with $\text{rank}(\boldsymbol{M}) = r$, and $\boldsymbol{m}$ a vector of order r with specified constants such that $\text{rank}(\boldsymbol{M}^T, \boldsymbol{m}) = r$. If we are interested in testing

$$H_0 : \boldsymbol{M}^T \boldsymbol{\beta} = \boldsymbol{m}, \tag{1.3}$$

the Wald test statistic is usually used in which $\boldsymbol{\beta}$ is estimated using the maximum likelihood estimator (MLE). Notice that if we consider $\boldsymbol{M} = \boldsymbol{I}_{k+1}$ and $\boldsymbol{m} = \boldsymbol{\beta}_0$, we get the Wald-type test statistic presented by Bianco and Martinez (2009) based on a weighted Bianco and Yohai (1996) estimator. It is well known that the MLE of $\boldsymbol{\beta}$ can be severely affected by outlying observations. Croux and Haesbroeck (2003) discuss the breakdown behavior of the MLE in the logistic regression model and show that the MLE breaks down when several outliers are added to a data set. In the recent years several authors have attempted to derive robust estimates of the parameters in the logistic regression model; see for instance Pregibon (1982), Morgenthaler (1992), Carroll and Pederson (1993), Christmann (1994), Bianco and Yohai (1996), Croux and Haesbroeck (2003), Bondell (2005; 2008) and Hobza et al. (2008; 2017). Our interest in this paper is to present a family of Wald-type test statistics based on the robust minimum density power divergence estimator for testing the general linear hypothesis given in (1.3).

In Section 2 we present the minimum density power divergence estimator for $\boldsymbol{\beta}$. The Wald-type test statistics, based on the minimum density power divergence estimator, are presented in Section 3, together with their asymptotic

properties. The theoretical robustness properties are presented in Section 4 and finally, Section 5 and 6 are devoted to the presentation of a simulation study and real data examples, respectively.

2. Minimum density power divergence estimator

If we denote by $y_{1},,...,y_n$ the observed values of the random variables $Y_1,...,Y_n$, the likelihood function for the logistic regression model is given by

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \{\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\}^{y_i} \left\{1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}^{1-y_i}. \tag{2.1}$$

So the MLE of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, is obtained by minimizing the log-likelihood function over $\boldsymbol{\beta}$ belonging to

$$\Theta = \left\{(\beta_0,...,\beta_k)^T : \beta_j \in (-\infty,\infty),\ j=0,...,k\right\} = \mathbb{R}^{k+1}.$$

We consider the probability vectors,

$$\widehat{\boldsymbol{p}} = \left(\frac{y_1}{n}, \frac{1-y_1}{n}, \frac{y_2}{n}, \frac{1-y_2}{n}, ..., \frac{y_n}{n}, \frac{1-y_n}{n}\right)^T$$

and

$$\boldsymbol{p}(\boldsymbol{\beta}) = \left(\pi(\boldsymbol{x}_1^T\boldsymbol{\beta})\frac{1}{n}, \left(1-\pi(\boldsymbol{x}_1^T\boldsymbol{\beta})\right)\frac{1}{n}, ..., \pi(\boldsymbol{x}_n^T\boldsymbol{\beta})\frac{1}{n}, \left(1-\pi(\boldsymbol{x}_n^T\boldsymbol{\beta})\right)\frac{1}{n}\right)^T.$$

The Kullback-Leibler divergence measure between the probability vectors $\widehat{\boldsymbol{p}}$ and $\boldsymbol{p}(\boldsymbol{\beta})$ is given by

$$d_{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) = \sum_{i=1}^{n}\sum_{j=1}^{2} \frac{y_{ij}}{n} \log \frac{y_{ij}}{\pi_j(\boldsymbol{x}_i^T\boldsymbol{\beta})}, \tag{2.2}$$

where

$$\pi_1(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \pi(\boldsymbol{x}_i^T\boldsymbol{\beta}),\ \pi_2(\boldsymbol{x}_i^T\boldsymbol{\beta}) = 1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta}),\ y_{i1} = y_i \text{ and } y_{i2} = 1-y_i.$$

It is not difficult to establish that

$$d_{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) = c - \frac{1}{n}\log \mathcal{L}(\boldsymbol{\beta}). \tag{2.3}$$

Here c is a constant independent of $\boldsymbol{\beta}$. Therefore, the MLE of $\boldsymbol{\beta}$ can be defined by

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\Theta} d_{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})). \tag{2.4}$$

Based on (2.4) we can use any divergence measure $d(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}))$ in order to define a minimum divergence estimator for $\boldsymbol{\beta}$. In this paper we shall use the density power divergence measure defined by Basu et al. (1998) because the minimum density power divergence estimators have excellent robustness properties, see for instance Basu et al. (2011; 2013; 2015; 2016), Ghosh et al. (2015; 2016b). The density power divergence between the probability vectors $\widehat{\boldsymbol{p}}$ and $\boldsymbol{p}(\boldsymbol{\beta})$ is given by

$$d_\lambda\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right)=\frac{1}{n^{1+\lambda}}\left\{\sum_{i=1}^{n}\left(\sum_{j=1}^{2}\pi_j^{1+\lambda}(\boldsymbol{x}_i^T\boldsymbol{\beta})-\left(1+\frac{1}{\lambda}\right)\sum_{j=1}^{2}y_{ij}\pi_j^{\lambda}(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)+\frac{n}{\lambda}\right\} \tag{2.5}$$

for $\lambda > 0$. For $\lambda = 0$, we have

$$d_0\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right)=\lim_{\lambda\to 0} d_\lambda\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right)=d_{KL}\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right).$$

Based on (2.4) and (2.5), we shall define the minimum density power divergence estimator as follows.

Definition 2.1. *The minimum density power divergence estimator for the parameter $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_\lambda$, in the logistic regression model is given by*

$$\widehat{\boldsymbol{\beta}}_\lambda=\arg\min_{\boldsymbol{\beta}\in\Theta} d_\lambda\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right),$$

where $d_\lambda\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right)$ is as defined in (2.5).

In order to obtain the estimating equations we need to get the derivative of (2.5) with respect to $\boldsymbol{\beta}$. First we write expression (2.5) as,

$$\begin{aligned} d_\lambda\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right)=\frac{1}{n^{1+\lambda}}\Bigg\{\sum_{i=1}^{n}\Big(&\pi^{1+\lambda}(\boldsymbol{x}_i^T\boldsymbol{\beta})+\left(1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)^{1+\lambda}\\ &-\left(1+\frac{1}{\lambda}\right)\left(y_i\pi^{\lambda}(\boldsymbol{x}_i^T\boldsymbol{\beta})+(1-y_i)\left(1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)^{\lambda}\right)\Big)+\frac{n}{\lambda}\Bigg\}. \end{aligned}$$

Now, taking into account the expressions

$$\begin{aligned} \frac{\partial\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}&=\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\left(1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)\boldsymbol{x}_i \text{ and}\\ \frac{\partial\left(1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)}{\partial\boldsymbol{\beta}}&=-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\left(1-\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})\right)\boldsymbol{x}_i \end{aligned}$$

and after some algebra, we get

$$\frac{\partial d_\lambda\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\boldsymbol{\beta}\right)\right)}{\partial\boldsymbol{\beta}}=\frac{1+\lambda}{n^{\lambda+1}}\sum_{i=1}^{n}(e^{\lambda\boldsymbol{x}_i^T\boldsymbol{\beta}}+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}-y_i(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})^{\lambda+2}}\boldsymbol{x}_i.$$

Therefore, the estimating equations for $\lambda > 0$ are given by

$$\sum_{i=1}^{n}\frac{e^{\lambda\boldsymbol{x}_i^T\boldsymbol{\beta}}+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})^{\lambda+1}}\left(\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})-y_i\right)\boldsymbol{x}_i=\boldsymbol{0}, \tag{2.6}$$

where $\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})$ as given in (1.2). Based on the previous results we have established the following theorem.

Theorem 2.1. *The minimum density power divergence estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_\lambda$, can be obtained as the solution of the system of equations given in (2.6).*

If we consider $\lambda = 0$ in (2.6), we get the estimating equations for the MLE as

$$\sum_{i=1}^{n}\left(\pi(\boldsymbol{x}_i^T\boldsymbol{\beta})-y_i\right)\boldsymbol{x}_i=\mathbf{0}.$$

Based on equation (2.6), we can write the estimating equation for the MDPDE under the the logistic regression model as

$$\sum_{i=1}^{n}\boldsymbol{\Psi}_\lambda\left(\boldsymbol{x}_i,y_i,\boldsymbol{\beta}\right)=\mathbf{0},$$

with

$$\boldsymbol{\Psi}_\lambda\left(\boldsymbol{x}_i,y_i,\boldsymbol{\beta}\right)=(e^{\lambda\boldsymbol{x}_i^T\boldsymbol{\beta}}+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}-y_i(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})^{\lambda+2}}\boldsymbol{x}_i. \tag{2.7}$$

In order to get the asymptotic distribution of the MDPDE of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_\lambda$, we are going to assume that not only are the explanatory variables random but they are also identically distributed and moreover

$$\left(\boldsymbol{X}_1,Y_1\right),....,\left(\boldsymbol{X}_n,Y_n\right)$$

are independent and identically distributed. We shall assume that $\boldsymbol{X}_1,...,\boldsymbol{X}_n$ is a random sample from a random variable $\boldsymbol{X}$ with marginal distribution function $H(\boldsymbol{x})$. By following the method given in Maronna et al. (2006), the asymptotic variance covariance matrix of $\sqrt{n}\,\widehat{\boldsymbol{\beta}}_\lambda$ is

$$\boldsymbol{J}_\lambda^{-1}\left(\boldsymbol{\beta}_0\right)\boldsymbol{K}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{J}_\lambda^{-1}\left(\boldsymbol{\beta}_0\right),$$

where

$$\begin{aligned}\boldsymbol{K}_\lambda\left(\boldsymbol{\beta}\right)&=E\left[\boldsymbol{\Psi}_\lambda\left(\boldsymbol{X},Y,\boldsymbol{\beta}\right)\boldsymbol{\Psi}_\lambda^T\left(\boldsymbol{X},Y,\boldsymbol{\beta}\right)\right]\\&=\int_{\mathcal{X}}E\left[\boldsymbol{\Psi}_\lambda\left(\boldsymbol{x},Y,\boldsymbol{\beta}\right)\boldsymbol{\Psi}_\lambda^T\left(\boldsymbol{x},Y,\boldsymbol{\beta}\right)\right]dH(\boldsymbol{x}),\end{aligned}$$

$\mathcal{X}$ is the support of $\boldsymbol{X}$, and

$$\boldsymbol{J}_\lambda\left(\boldsymbol{\beta}\right)=E\left[\frac{\partial\boldsymbol{\Psi}_\lambda\left(\boldsymbol{X},Y,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^T}\right]=\int_{\mathcal{X}}E\left[\frac{\partial\boldsymbol{\Psi}_\lambda\left(\boldsymbol{x},Y,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^T}\right]dH(\boldsymbol{x}).$$

In relation to the matrix $\boldsymbol{K}_\lambda\left(\boldsymbol{\beta}_0\right)$, we have

$$\begin{aligned}&E\left[\boldsymbol{\Psi}_\lambda\left(\boldsymbol{x},Y,\boldsymbol{\beta}\right)\boldsymbol{\Psi}_\lambda^T\left(\boldsymbol{x},Y,\boldsymbol{\beta}\right)\right]\\&=\frac{(e^{\lambda\boldsymbol{x}^T\boldsymbol{\beta}}+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^2}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{2(\lambda+2)}}E\left[\left(e^{\boldsymbol{x}^T\boldsymbol{\beta}}-Y(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})\right)^2\right]\boldsymbol{x}\boldsymbol{x}^T,\end{aligned}$$

but $E\left[Y^2\right]=\pi(\boldsymbol{x}^T\boldsymbol{\beta})$ and

$$E\left[\left(e^{\boldsymbol{x}^T\boldsymbol{\beta}}-Y(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})\right)^2\right]=e^{\boldsymbol{x}^T\boldsymbol{\beta}}.$$

Therefore

$$\boldsymbol{K}_\lambda\left(\boldsymbol{\beta}\right)=E\left[\boldsymbol{\Psi}_\lambda\left(\boldsymbol{X},Y,\boldsymbol{\beta}\right)\boldsymbol{\Psi}_\lambda^T\left(\boldsymbol{X},Y,\boldsymbol{\beta}\right)\right]=\int_{\mathcal{X}}\frac{(e^{\lambda\boldsymbol{x}^T\boldsymbol{\beta}}+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^2}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{2(\lambda+2)}}e^{\boldsymbol{x}^T\boldsymbol{\beta}}\boldsymbol{x}\boldsymbol{x}^T dH(\boldsymbol{x}). \tag{2.8}$$

An estimator of $\boldsymbol{K}_\lambda(\boldsymbol{\beta})$ will be

$$\widehat{\boldsymbol{K}}_\lambda(\boldsymbol{\beta}) = \int_{\mathcal{X}} \frac{(e^{\lambda \boldsymbol{x}^T\boldsymbol{\beta}} + e^{\boldsymbol{x}^T\boldsymbol{\beta}})^2}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{2(\lambda+2)}} e^{\boldsymbol{x}^T\boldsymbol{\beta}} \boldsymbol{x}\boldsymbol{x}^T dH_n(\boldsymbol{x}),$$

where $H_n(\boldsymbol{x})$ the empirical distribution function associated with the sample $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Then

$$\widehat{\boldsymbol{K}}_\lambda(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{(e^{\lambda \boldsymbol{x}^T\boldsymbol{\beta}} + e^{\boldsymbol{x}^T\boldsymbol{\beta}})^2}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{2(\lambda+2)}} e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \boldsymbol{x}_i\boldsymbol{x}_i^T. \tag{2.9}$$

It is interesting to observe that for $\lambda = 0$ we get

$$\begin{aligned}
\widehat{\boldsymbol{K}}_0(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})^2}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})^4} e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \boldsymbol{x}_i\boldsymbol{x}_i^T \\
&= \frac{1}{n}\mathbb{X}^T \operatorname{diag}\left(\pi_i(\boldsymbol{x}^T\boldsymbol{\beta})\left(1-\pi_i(\boldsymbol{x}^T\boldsymbol{\beta})\right)\right)_{i=1,\ldots,n} \mathbb{X} \\
&= \boldsymbol{I}_F(\boldsymbol{\beta}),
\end{aligned}$$

with $\boldsymbol{I}_F(\boldsymbol{\beta})$ being the Fisher information matrix associated to the logistic regression model.

To compute the matrix $\boldsymbol{J}_\lambda(\boldsymbol{\beta})$, first we need to calculate

$$\frac{\partial \boldsymbol{\Psi}_\lambda(\boldsymbol{x}, y, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = L_1(\boldsymbol{x}, y, \boldsymbol{\beta}) + L_2(\boldsymbol{x}, y, \boldsymbol{\beta}),$$

where

$$L_1(\boldsymbol{x}, y, \boldsymbol{\beta}) = (\lambda e^{\lambda \boldsymbol{x}^T\boldsymbol{\beta}} + e^{\boldsymbol{x}^T\boldsymbol{\beta}}) \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}} - y(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{\lambda+2}} \boldsymbol{x}\boldsymbol{x}^T$$

and

$$\begin{aligned}
L_2(\boldsymbol{x}, y, \boldsymbol{\beta}) &= (e^{\lambda \boldsymbol{x}^T\boldsymbol{\beta}} + e^{\boldsymbol{x}^T\boldsymbol{\beta}}) \\
&\quad \times \left(\frac{\left(e^{\boldsymbol{x}^T\boldsymbol{\beta}} - y e^{\boldsymbol{x}^T\boldsymbol{\beta}}\right)}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{\lambda+2}} - \frac{(\lambda+2)\, e^{\boldsymbol{x}^T\boldsymbol{\beta}} \left(e^{\boldsymbol{x}^T\boldsymbol{\beta}} - y(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})\right)}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{\lambda+3}} \right) \boldsymbol{x}\boldsymbol{x}^T,
\end{aligned}$$

and hence

$$E\left[\frac{\partial \boldsymbol{\Psi}_\lambda(\boldsymbol{x}, Y, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right] = E\left[L_1(\boldsymbol{x}, Y, \boldsymbol{\beta})\right] + E\left[L_2(\boldsymbol{x}, Y, \boldsymbol{\beta})\right].$$

But

$$E\left[e^{\boldsymbol{x}^T\boldsymbol{\beta}} - Y(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})\right] = e^{\boldsymbol{x}^T\boldsymbol{\beta}} - \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}^T\boldsymbol{\beta}}}(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}}) = 0.$$

Therefore

$$E\left[L_1(\boldsymbol{x}, Y, \boldsymbol{\beta})\right] = \mathbf{0}_{(k+1)(k+1)}.$$

On the other hand

$$\begin{aligned} E\left[L_2(\boldsymbol{x},Y,\boldsymbol{\beta})\right] &= \frac{e^{\lambda \boldsymbol{x}^T\boldsymbol{\beta}}+e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{2(\lambda+2)}}\left((1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{\lambda+2}E\left[e^{\boldsymbol{x}^T\boldsymbol{\beta}}-Ye^{\boldsymbol{x}^T\boldsymbol{\beta}}\right]\right.\\ &\quad\left.+(\lambda+2)\,(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{\lambda+1}e^{\boldsymbol{x}^T\boldsymbol{\beta}}E\left[e^{\boldsymbol{x}^T\boldsymbol{\beta}}-Y(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})\right]\right)\boldsymbol{x}\boldsymbol{x}^T\\ &= \frac{e^{\lambda \boldsymbol{x}^T\boldsymbol{\beta}}+e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{\lambda+3}}e^{\boldsymbol{x}^T\boldsymbol{\beta}}\boldsymbol{x}\boldsymbol{x}^T. \end{aligned}$$

Finally,

$$\begin{aligned} \boldsymbol{J}_\lambda(\boldsymbol{\beta}) &= \int_{\mathcal{X}} E\left[\frac{\partial \boldsymbol{\Psi}_\lambda(\boldsymbol{x},Y,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]dH(\boldsymbol{x}) \qquad (2.10)\\ &= \int_{\mathcal{X}} \frac{e^{\lambda \boldsymbol{x}^T\boldsymbol{\beta}}+e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{\lambda+3}}e^{\boldsymbol{x}^T\boldsymbol{\beta}}\boldsymbol{x}\boldsymbol{x}^T dH(\boldsymbol{x}), \end{aligned}$$

and an estimator of $\boldsymbol{J}_\lambda(\boldsymbol{\beta})$ is given by

$$\widehat{\boldsymbol{J}}_\lambda(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{\lambda \boldsymbol{x}_i^T\boldsymbol{\beta}}+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}})^{\lambda+3}}e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\boldsymbol{x}_i\boldsymbol{x}_i^T. \qquad (2.11)$$

In particular, for $\lambda = 0$, we have

$$\begin{aligned} \widehat{\boldsymbol{J}}_0(\boldsymbol{\beta}) &= \frac{1}{n}\mathbb{X}^T\ \mathrm{diag}\left(\pi_i(\boldsymbol{x}^T\boldsymbol{\beta})\left(1-\pi_i(\boldsymbol{x}^T\boldsymbol{\beta})\right)\right)_{i=1,\ldots,n}\mathbb{X}\\ &= \boldsymbol{I}_F(\boldsymbol{\beta}). \end{aligned}$$

From the sequence of above results, the next theorem follows.

Theorem 2.2. *The asymptotic distribution of the MDPDE for $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_\lambda$, is given by*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}_0)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(\boldsymbol{0},\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\right)$$

where

$$\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0) = \boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{K}_\lambda(\boldsymbol{\beta}_0)\,\boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)$$

and the matrices $\boldsymbol{J}_\lambda(\boldsymbol{\beta}_0)$ and $\boldsymbol{K}_\lambda(\boldsymbol{\beta}_0)$ where defined in (2.10) and (2.8), respectively.

Remark 2.1. *We have considered that the covariates are random, a crucial assumption to get the asymptotic distribution of the MDPDE by using the standard asymptotic theory for M-estimators. It is interesting to highlight that whenever the covariates were non-stochastic (fixed design case), the asymptotic distribution of the MDPDE could be obtained from Ghosh et al. (2016d) without using the standard asymptotic theory of M-estimators. In order to present the results in the most general setting, we shall assume that the random variables Y_i with $i = 1,...,I$, are binomial with parameters n_i and $\pi_i = \pi(\boldsymbol{x}_i^T\boldsymbol{\beta})$ instead of Bernoulli random variables. We shall denote by $N = \sum_{i=1}^{I} n_i$ and let n_{i1} denotes the observed value of Y_i. We will assume that I is fixed and for each $i = 1,\ldots,I$, construct the independent and identically distributed latent observations $z_{i1},\ldots,z_{in_i}$ each following a Bernoulli distribution with probability π*

and $n_{i1} = \sum_{j=1}^{n_i} z_{ij}$. *Then,* N *random observations* $z_{11}, \ldots, z_{1n_1}$, $z_{21}, \ldots, z_{2n_2}$, $\ldots$, $z_{I1}, \ldots, z_{In_I}$ *are independent but have possibly different distribution with* $z_{ij} \sim Ber(\pi_i)$. *This falls under the general setup of independent but non-homogeneous observations as considered in Ghosh and Basu (2013) and hence it is immediately seen that the corresponding estimating equations for the MDPDE,* $\widehat{\boldsymbol{\beta}}^*_\lambda$ *in this context, for* $\lambda > 0$ *are given by*

$$\sum_{i=1}^{I} \frac{e^{\lambda \boldsymbol{x}_i^T \boldsymbol{\beta}} + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})^{\lambda+1}} \left(n_i \pi(\boldsymbol{x}_i^T \boldsymbol{\beta}) - n_{i1}\right) \boldsymbol{x}_i = \mathbf{0}$$

and for $\lambda = 0$, *by*

$$\sum_{i=1}^{I} \left(n_i \pi(\boldsymbol{x}_i^T \boldsymbol{\beta}) - n_{i1}\right) \boldsymbol{x}_i = \mathbf{0}. \tag{2.12}$$

Now, assuming

$$\lim_{N\to\infty} \frac{n_i}{N} = \alpha_i \in (0,1)\,, \;\; i = 1, ..., I,$$

and following Ghosh and Basu (2013), we get the asymptotic distribution of the MDPDE of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}^*_\lambda$, *as given by*

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}^*_\lambda - \boldsymbol{\beta}_0) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^*\left(\boldsymbol{\beta}_0\right)\right) \tag{2.13}$$

where

$$\boldsymbol{\Sigma}^*\left(\boldsymbol{\beta}_0\right) = \boldsymbol{J}^{*-1}\left(\boldsymbol{\beta}_0\right) \boldsymbol{K}^*\left(\boldsymbol{\beta}_0\right) \boldsymbol{J}^{*-1}\left(\boldsymbol{\beta}_0\right).$$

Here, the matrices $\boldsymbol{J}^*\left(\boldsymbol{\beta}_0\right)$ *and* $\boldsymbol{K}^*\left(\boldsymbol{\beta}_0\right)$ *can be obtained directly from the general results of Ghosh and Basu (2013) or from the simplified results in the context of Bernoulli logistic regression with fixed design in Ghosh and Basu (2015) and are given by*

$$\boldsymbol{J}^*\left(\boldsymbol{\beta}_0\right) = \sum_{i=1}^{I} \alpha_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} \frac{e^{\lambda \boldsymbol{x}_i^T \boldsymbol{\beta}} + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})^{\lambda+3}} \boldsymbol{x}_i \boldsymbol{x}_i^T,$$

and

$$\boldsymbol{K}^*\left(\boldsymbol{\beta}_0\right) = \sum_{i=1}^{I} \alpha_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} \frac{(e^{\lambda \boldsymbol{x}_i^T \boldsymbol{\beta}} + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})^2}{(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})^{2(\lambda+2)}} \boldsymbol{x}_i \boldsymbol{x}_i^T.$$

For $\lambda = 0$, *it is clear, based on (2.12), that we get the classical likelihood estimator. We can observe that in this situation*

$$\boldsymbol{J}^*\left(\boldsymbol{\beta}_0\right) = \boldsymbol{K}^*\left(\boldsymbol{\beta}_0\right) = \boldsymbol{I}_F\left(\boldsymbol{\beta}_0\right)$$

and we get the classical result,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}^*_{\lambda=0} - \boldsymbol{\beta}_0) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_F^{-1}\left(\boldsymbol{\beta}_0\right)\right).$$

3. Wald type test statistic for testing linear hypothesis

Based on the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_\lambda$ we are going to define a family of Wald-type test statistics for testing the null hypothesis

$$H_0 : \boldsymbol{M}^T\boldsymbol{\beta} = \boldsymbol{m}, \tag{3.1}$$

where $\boldsymbol{M}^T$ is any matrix of r rows and $k+1$ columns and m a vector of order r of specified constant. We assume that the matrix $\boldsymbol{M}^T$ has full row rank, i.e., $\text{rank}\,(\boldsymbol{M}) = r$.

Definition 3.1. *Let $\widehat{\boldsymbol{\beta}}_\lambda$ be the minimum power divergence estimator. The family of Wald type test statistics for testing the null hypothesis given in (3.1) is given by*

$$\begin{aligned} W_n &= n(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{m})^T \left(\boldsymbol{M}^T\boldsymbol{J}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\lambda)\boldsymbol{K}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\boldsymbol{J}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\lambda)\boldsymbol{M}\right)^{-1}(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{m}) \\ &= n(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{m})^T(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\boldsymbol{M})^{-1}(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{m}). \end{aligned} \tag{3.2}$$

In the particular case of $\lambda = 0$, i.e. $\widehat{\boldsymbol{\beta}}$ is the MLE, we get the classical Wald test statistic because in this case

$$\boldsymbol{J}_{\lambda=0}^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{K}_{\lambda=0}(\boldsymbol{\beta}_0)\,\boldsymbol{J}_{\lambda=0}^{-1}(\boldsymbol{\beta}_0) = \boldsymbol{I}_F^{-1}(\boldsymbol{\beta}_0).$$

Theorem 3.1. *The asymptotic distribution of the Wald type test statistic, W_n, defined in (3.2), under the null hypothesis given in (3.1), is a chi-square distribution with r degrees of freedom.*

Proof. We have $\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{m} = \boldsymbol{M}^T(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0)$ and $\sqrt{n}(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0))$. Therefore

$$\sqrt{n}(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{m}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\,\boldsymbol{M}\right)$$

and since $\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\,\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{K}_\lambda(\boldsymbol{\beta}_0)\,\boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{M}\right)^{-1} = \boldsymbol{I}_{r\times r}$, the asymptotic distribution of W_n is a chi-square distribution with r degrees of freedom. □

Remark 3.1. *If we consider*

$$\boldsymbol{M}^T = \begin{pmatrix} \boldsymbol{0}_{k\times 1} & \boldsymbol{I}_{k\times k} \end{pmatrix}_{k\times(k+1)} \tag{3.3}$$

we have

$$\boldsymbol{M}^T\boldsymbol{\beta} = \boldsymbol{0},$$

if and only if $\beta_i = 0$, $i = 1, ..., k$. Therefore, we can consider the Wald-type test statistics with $\boldsymbol{M}^T$ defined in (3.3) for testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

In this case, the asymptotic distribution of the Wald type test statistic is a chi square distribution with k degrees of freedom. If we consider $\boldsymbol{M}^T$ to be a vector with all elements equal zero except for the $(i+1)$-th term, equals 1, we can test

$$H_0 : \beta_i = 0.$$

Based on the previous theorem the null hypothesis given in (3.1) will be rejected if we have that

$$W_n > \chi^2_{r,\alpha}, \tag{3.4}$$

where $\chi^2_{r,\alpha}$ is the quantile of order $1-\alpha$.for a chi-square with r degrees of freedom Let us consider $\boldsymbol{\beta}^* \in \Theta$ such that $\boldsymbol{M}^T\boldsymbol{\beta}^* \neq \boldsymbol{m}$, i.e., $\boldsymbol{\beta}^*$ does not belong to the null hypothesis. We denote

$$q_{\boldsymbol{\beta}_1}(\boldsymbol{\beta}_2) = \left(\boldsymbol{M}^T\boldsymbol{\beta}_1 - \boldsymbol{m}\right)^T \left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_2)\,\boldsymbol{M}\right)^{-1} \left(\boldsymbol{M}^T\boldsymbol{\beta}_1 - \boldsymbol{m}\right)$$

and we are going to get an approximation to the power function for the test statistics given in (3.4).

Theorem 3.2. *Let $\boldsymbol{\beta}^* \in \Theta$, with $\boldsymbol{M}^T\boldsymbol{\beta}^* \neq \boldsymbol{m}$, be the true value of the parameter so that $\widehat{\boldsymbol{\beta}}_\lambda \xrightarrow[n\to\infty]{P} \boldsymbol{\beta}^*$. The power function of the test statistic given in (3.4), in $\boldsymbol{\beta}^*$, is given by*

$$\xi(\boldsymbol{\beta}^*) = 1 - \Phi_n\left(\frac{1}{\sigma(\boldsymbol{\beta}^*)}\left(\frac{\chi^2_{r,\alpha}}{\sqrt{n}} - \sqrt{n}q_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)\right), \tag{3.5}$$

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

$$\sigma^2(\boldsymbol{\beta}^*) = \left.\frac{\partial q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0) \left.\frac{\partial q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

Proof. We have

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

Now we are going to get the asymptotic distribution of the random variable $\sqrt{n}(q_{\widehat{\boldsymbol{\beta}}_\lambda}(\widehat{\boldsymbol{\beta}}_\lambda) - q_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*))$. It is clear that $q_{\widehat{\boldsymbol{\beta}}_\lambda}(\widehat{\boldsymbol{\beta}}_\lambda)$ and $q_{\widehat{\boldsymbol{\beta}}_\lambda}(\boldsymbol{\beta}^*)$ have the same asymptotic distribution because $\widehat{\boldsymbol{\beta}}_\lambda \xrightarrow[n\to\infty]{P} \boldsymbol{\beta}^*$. A first order Taylor expansion of $q_{\widehat{\boldsymbol{\beta}}_\lambda}(\boldsymbol{\beta}^*)$ at $\widehat{\boldsymbol{\beta}}_\lambda$ around $\boldsymbol{\beta}^*$ gives

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

Therefore it holds

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

where

$$\sigma^2(\boldsymbol{\beta}^*) = \left.\frac{\partial q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{K}_\lambda(\boldsymbol{\beta}_0)\,\boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0) \left.\frac{\partial q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

Now the result follows. □

Remark 3.2. *Based on the previous theorem we can obtain the sample size necessary to get a fix power* $\xi(\boldsymbol{\beta}^*) = \xi_0$. *From (3.5), we must solve the equation*

$$1 - \xi_0 = \Phi\left(\frac{1}{\sigma(\boldsymbol{\beta}^*)}\left(\frac{\chi^2_{r,\alpha}}{\sqrt{n}} - \sqrt{n} q_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\right)\right)$$

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

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

being

$$A = \sigma^2(\boldsymbol{\beta}^*)\left(\Phi^{-1}(1-\xi_0)\right)^2 \text{ and } B = 2q_{\boldsymbol{\beta}^*}(\boldsymbol{\beta}^*)\chi^2_{r,\alpha}.$$

In the following theorem we present an approximation to the power function at the contiguous alternative hypothesis

$$\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{d}, \tag{3.6}$$

with $\boldsymbol{d}$ satisfying $\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{d} \in \Theta$.

Theorem 3.3. *An approximation of the power function for the test statistic given in (3.4), in* $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{d}$ *is given by*

$$\xi(\boldsymbol{\beta}_n) = 1 - F_{\chi^2_r(\delta)}\left(\chi^2_{r,\alpha}\right),$$

where $F_{\chi^2_r(\delta)}$ *is the distribution function of a non-central chi-square with* p *degrees of freedom and non-centrality parameter* δ *given by* $\delta = \boldsymbol{d}^T \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{d}$.

4. Robustness analysis

4.1. Influence function of the MDPDE

We will consider the influence function analysis of Hampel et al. (1986) to study the robustness of our proposed MDPDE and the corresponding Wald-type test of general linear hypothesis in the logistic regression model. Since the MDPDE can be written in term of a M-estimator as shown in Section 2 with ψ-function given by (2.7), we can apply directly the results of the M-estimation theory of Hampel et al. (1986) in order to get the influence function of the proposed MDPDE.

However, we first need to re-define the minimum density power divergence estimator $\widehat{\boldsymbol{\beta}}_\lambda$ from Definition 1 in terms of a statistical functional. Let us assume the stochastic nature of the covariates X and that the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with some joint distribution G. Then we define the required statistical functional corresponding to $\widehat{\boldsymbol{\beta}}_\lambda$ as follows.

Definition 4.1. *The minimum DPD functional $T_\lambda(G)$, corresponding to the minimum DPD estimator $\widehat{\boldsymbol{\beta}}_\lambda$, at the joint distribution G is defined as the solution of the system of equations*

$$E_G\left[\boldsymbol{\Psi}_\lambda(\boldsymbol{X}, Y, \boldsymbol{\beta})\right] = \mathbf{0}$$

with respect to $\boldsymbol{\beta}$, whenever the solution exists.

Now, if G_0 denotes the joint model distribution with the true parameter value $\boldsymbol{\beta}_0$ under which

$$P_{G_0}(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i^T \boldsymbol{\beta}_0),$$

then it is easy to see that $E_{G_0}\left[\Psi_\lambda(\boldsymbol{X}, Y, \boldsymbol{\beta}_0)\right] = 0$ and hence $\boldsymbol{T}_\lambda(G_0) = \boldsymbol{\beta}_0$. Therefore, the minimum DPD functional $\boldsymbol{T}_\lambda$ is Fisher consistent.

Next, we can easily obtain the influence function for our MDPDE at the model distribution G_0 as presented in the following theorem. This can be derived either through a straightforward calculation or by applying the corresponding results from M-estimation theory of Hampel et al. (1986) and hence the proof of the theorem is omitted.

Theorem 4.1. *The influence function of the minimum DPD functional T_λ, as defined in Definition 4.1 with tuning parameter λ, at the model distribution G_0 is given by*

$$\begin{aligned} \mathcal{IF}((\boldsymbol{x}_t, y_t), T_\lambda, G_0) &= \boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\left(\boldsymbol{\Psi}_\lambda(\boldsymbol{x}_t, y_t, \boldsymbol{\beta}_0) - E_{G_0}[\boldsymbol{\Psi}_\lambda(\boldsymbol{X}, Y, \boldsymbol{\beta}_0)]\right) \\ &= \boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\boldsymbol{\Psi}_\lambda(\boldsymbol{x}_t, y_t, \boldsymbol{\beta}_0), \end{aligned}$$

where $\boldsymbol{J}_\lambda(\boldsymbol{\beta})$ is as defined in Section 2 of the paper and $(\boldsymbol{x}_t, y_t)$ is the point of contamination.

Before studying the above influence function, let us first recall different types of outliers in logistic regression model following the discussion in Croux and Haesbroeck (2003). A contamination point (x_t, y_t) will be a leverage point if x_t is outlying in the covariates space and will be a vertical outlier (in response) if it is not a leverage point but the residual $y_t - \pi(\boldsymbol{x}_t^T \boldsymbol{\beta})$ is large. Croux and Haesbroeck (2003) also noted that, for the maximum likelihood estimator of $\boldsymbol{\beta}$, a vertical outlier or a "good" leverage point (for which the residual is small) has bounded influence whereas a bad leverage point (e.g., misclassified observation etc.) has infinite influence for $||\boldsymbol{x}_t|| \to \infty$.

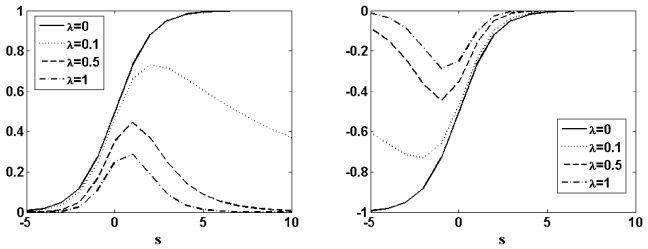
Next, in order to study the similar nature of the influence function of the MDPDE having different λ, note that the influence function given in Theorem 4.1 can be factored into two components as

$$\mathcal{IF}((\boldsymbol{x}_t, y_t), T_\lambda, G_0) = \widetilde{\Psi}_\lambda(\boldsymbol{x}_t^T \boldsymbol{\beta}_0, y_t) \boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0) \boldsymbol{x}_t,$$

where the first part $\widetilde{\boldsymbol{\Psi}}_\lambda$ depends on the score, $s = \boldsymbol{x}_t^T \boldsymbol{\beta}_0$, and the response, y_t, and is defined as

$$\widetilde{\Psi}_\lambda(s, y) = \frac{\left(e^{\lambda s} + e^s\right)\left(e^s - y(1 + e^s)\right)}{\left(1 + e^s\right)^{\lambda+2}}.$$

Figure 1 shows the nature of this function over the score input at $y = 0, 1$ for different values of λ. Clearly, the function $\widetilde{\Psi}_\lambda$ corresponding to $\lambda = 0$ (MLE) is unbounded as $s \to \infty$, illustrating the well-known non-robust nature of the MLE. However, for $\lambda > 0$ the function $\widetilde{\Psi}_\lambda$ is bounded in s and becomes more re-descending as λ increase, which implies the increasing robustness of our proposed MDPDEs with increasing $\lambda > 0$.



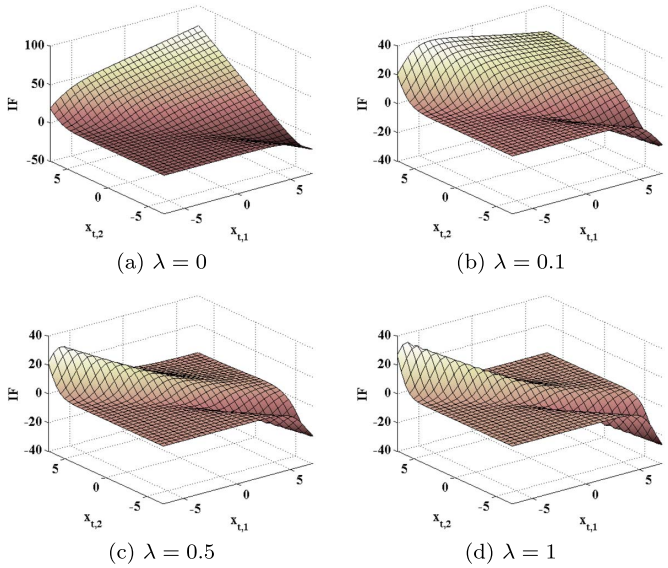
(a) $y_t = 0$ (b) $y_t = 1$

FIG 1. *Plots of* $\widetilde{\Psi}_\lambda(s; y)$ *over* s *for different* λ *and* $y = 0, 1$.

Further, to examine the effect of different types of leverage points more clearly, following Croux and Haesbroeck (2003), in Figure 2, we present the influence function of the MDPDE of the first slope parameter $\boldsymbol{\beta}_1$ over the covariates values in a logistic regression model with two independent standard normal covariates and $\boldsymbol{\beta}_0 = (0, 1, 1)^T$ fixing $y_t = 0$ (without loss of generality). We can see that when both covariates tend to $-\infty$ the influence function becomes zero for all MDPDEs including the MLE (at $\lambda = 0$). These are the "good" leverage points, as noted in Croux and Haesbroeck (2003), and all MDPDEs are robust with respect to such good leverages as in the case of MLE. However, when the covariates approach to ∞ they yield bad leverage points (generally corresponding to misclassified points) and have large influence for the MLE ($\lambda = 0$). But in this case the influence function values of the MDPDEs with $\lambda > 0$ are quite small even for these bad leverages and get progressively smaller as λ increases. This phenomenon gain indicates the increasing robustness of our proposed MDPDEs with larger positive λ.

Remark 4.1. *Under the setup of Remark 2.1, even when the covariates are non-stochastic, we can derive the influence function of the corresponding MDPDE,* $\widehat{\boldsymbol{\beta}}^*_\lambda$, *following Ghosh and Basu (2013). Whenever the covariates* $\boldsymbol{x}_i$*s are fixed, the contamination needs to be considered over the conditional distribution of the response given the covariates which are not identical for each groups with given fixed covariates. Hence, as in Ghosh and Basu (2013), we can consider the contamination in any one group or in all the group. This leads to the influence function of* $\widehat{\boldsymbol{\beta}}^*_\lambda$ *under contamination only in one group (*i_0*-th, say) with covariate* $\boldsymbol{x}_{i_0}$ *as given by*

$$\mathcal{IF}_{i_0}(y_{t_{i_0}}, T_\lambda, G_0) = \boldsymbol{J}^{*-1}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\Psi}_\lambda(\boldsymbol{x}_{i_0}, y_{t_{i_0}}, \boldsymbol{\beta}_0),$$



(a) $\lambda = 0$ (b) $\lambda = 0.1$

(c) $\lambda = 0.5$ (d) $\lambda = 1$

FIG 2. *Influence function of the MDPDE of the first slope parameter* β_1 *for different* λ *(*$y_t = 0$*).*

where $y_{t_{i_0}}$ *is the contamination point in the contaminated distribution of* Y *given* $\boldsymbol{X} = \boldsymbol{x}_{i_0}$*. Similarly, if there is contamination in all the groups with covariates* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I$*, respectively, at the contamination points* $y_{t_1}, \ldots, y_{t_I}$*, then the resulting influence function has the form*

$$\mathcal{IF}((y_{t_1}, \ldots, y_{t_I}), T_\lambda, G_0) = \boldsymbol{J}_\lambda^{*-1}(\boldsymbol{\beta}_0) \sum_{i=1}^{I} \boldsymbol{\Psi}_\lambda(\boldsymbol{x}_i, y_{t_i}, \boldsymbol{\beta}_0),$$

Note that, since the response in a logistic regression takes only values 0 *and* 1*, the contamination points* y_{t_i} *all take values only in* $\{0, 1\}$ *(misclassification errors) and hence all the above influence functions are bounded with respect to contamination in response for all* $\lambda \geq 0$*. Hence, the effect of these (misclassification) error in response cannot be clearly inferred only from these influence functions; see Pregibon (1982), Copas (1988) and Victoria-Feser (2000) for more examples such analysis of misclassification error in logistic regression with a fixed design. However, the above influence functions are bounded in the values of given fixed covariates only for* $\lambda > 0$*, implying the robustness of the MDPDEs with* $\lambda > 0$ *and non-robust nature of MLE (at* $\lambda = 0$*) with respect to the extreme values of the fixed design in any one group.*

4.2. Influence function of the Wald-Type test statistics

We will now study the robustness of the proposed Wald-type test of Section 3 through the influence function of the corresponding test statistics W_n defined

in Definition 5. Ignoring the multiplier n, let us define the associated statistical functional for the test statistics W_n evaluated at any joint distribution G as given by

$$W_\lambda(G) = \left(\boldsymbol{M}^T\boldsymbol{T}_\lambda(G) - \boldsymbol{m}\right)^T (\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\boldsymbol{M})^{-1} \left(\boldsymbol{M}^T\boldsymbol{T}_\lambda(G) - \boldsymbol{m}\right). \quad (4.1)$$

Now, considering the ε-contaminated joint distribution $G_\varepsilon = (1-\varepsilon)G + \varepsilon\wedge_{\boldsymbol{w}}$ with respect to the point mass contamination distribution $\wedge_{\boldsymbol{w}}$ at the contamination point $\boldsymbol{w} = (\boldsymbol{x}_t, y_t)$, the influence function of $W_\lambda(\cdot)$ is defined as

$$\begin{aligned}\mathcal{IF}(\boldsymbol{w}, W_\lambda, G) &= \left.\frac{\partial W_\lambda(G_\varepsilon)}{\partial\varepsilon}\right|_{\varepsilon=0} \\ &= \left(\boldsymbol{M}^T T_\lambda(G) - \boldsymbol{m}\right)^T \left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1} \boldsymbol{M}^T\mathcal{IF}(\boldsymbol{w}, T_\lambda, G).\end{aligned}$$

Now, assuming the null hypothesis to be true, let G_0 denote the joint model distribution with true parameter value $\boldsymbol{\beta}_0$ satisfying $\boldsymbol{M}^T\boldsymbol{\beta}_0 = \boldsymbol{m}$. Then, under G_0, we have $\boldsymbol{T}_\lambda(G_0) = \boldsymbol{\beta}_0$ and hence $\mathcal{IF}(\boldsymbol{w}, W_\lambda, G_0) = \boldsymbol{0}$. Therefore, the first order influence function analysis is not adequate to quantify the robustness of the proposed Wald-type test statistics W_λ. It is bounded in the contamination points $\boldsymbol{w} = (\boldsymbol{x}_t, y_t)$ for all $\lambda \geq 0$ but does not necessarily imply the robustness of the tests since it includes the well-known non-robust MLE based Wald-test at $\lambda = 0$. This fact is consistent with the robustness analysis of different other Wald-type tests under different setups (See, for example, Rousseeuw and Ronchetti, 1979; Toma and Broniatowski, 2011; Ghosh et al., 2016b etc.) and we need to consider the second order influence analysis to asses the robustness of W_λ.

The second order influence function of the Wald-type test statistics W_n at the joint distribution G is defined as

$$\begin{aligned}\mathcal{IF}_2(\boldsymbol{w}, W_\lambda, G) &= \left.\frac{\partial^2 W_\lambda(G_\varepsilon)}{\partial\varepsilon^2}\right|_{\varepsilon=0} \\ &= \left(\boldsymbol{M}^T\boldsymbol{T}_\lambda(G) - \boldsymbol{m}\right)^T \left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta})\boldsymbol{M}\right)^{-1} \boldsymbol{M}^T\mathcal{IF}_2(\boldsymbol{w}, \boldsymbol{T}_\lambda, G) \\ &+ \mathcal{IF}^T(\boldsymbol{w}, \boldsymbol{T}_\lambda, G)\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta})\boldsymbol{M}\right)^{-1} \boldsymbol{M}^T\mathcal{IF}(\boldsymbol{w}, \boldsymbol{T}_\lambda, G).\end{aligned}$$

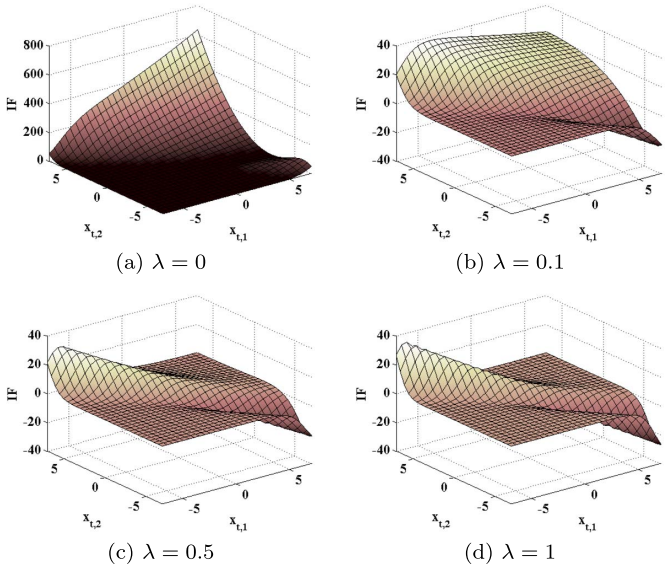
Again, under the null hypothesis H_0 with $\boldsymbol{\beta}_0$ being the corresponding true parameter value, this second order influence function simplifies further as presented in the following theorem and yields the possibility of studyng the robustness of our proposed tests through its boundedness.

Theorem 4.2. *The second order influence function of the proposed Wald-type test statistics W_n, given in Definition 5, at the null model distribution G_0 having true parameter value $\boldsymbol{\beta}_0$ is given by*

$$\begin{aligned}&\mathcal{IF}_2(\boldsymbol{w}, W_\lambda, G_0) \\ &= \mathcal{IF}^T(\boldsymbol{w}, \boldsymbol{T}_\lambda, G_0)\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1} \boldsymbol{M}^T\mathcal{IF}(\boldsymbol{w}, \boldsymbol{T}_\lambda, G_0).\end{aligned}$$

$$= \widetilde{\Psi}_\lambda^2(\boldsymbol{x}_t^T\boldsymbol{\beta}_0, y_t)\boldsymbol{x}_t^T\boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_0)\boldsymbol{x}_t.$$

Note that, the influence function of the Wald-type test statistic is directly a quadratic function of the corresponding MDPDE used. Hence, as described in the previous subsection, the influence function for the proposed tests with $\lambda > 0$ will be small and bounded for all kinds of outliers in a logistic regression model, whereas the classical MLE based Wald-type test will have an unbounded influence function for large "bad" leverage points. Figure 3 shows the plots of these second order influence functions for the Wald-type test statistics for different λ for testing the significance of the first slope parameter in a logistic regression model with two independent standard normal covariates and $\boldsymbol{\beta}_0 = (0, 1, 1)^T$ fixing $y_t = 0$. The behavior of the influence functions are again similar to those observed for the corresponding $MDPDE$ in Figure 3, which shows the greater robustness of our proposal at larger positive λ over the non-robust MLE based Wald test at $\lambda = 0$.



(a) $\lambda = 0$ (b) $\lambda = 0.1$

(c) $\lambda = 0.5$ (d) $\lambda = 1$

FIG 3. *Second order Influence function of the Wald-type test statistics for testing significance of the first slope parameter β_1 for different λ ($y_t = 0$).*

4.3. Level and power influence functions

We now study the robustness of the proposed tests through the stability of their Type-I and Type-II error which are the two basic components for measuring the performance of any testing procedure. In particular, we will study the local stability of level and power of the proposed tests through the corresponding influence function analysis. Note that the finite sample level and power of

our proposed Wald-type tests are difficult to compute and has no general form; on the other hand, the tests are consistent having asymptotic power equal one against any fixed alternative. So, we will study the influence function of the asymptotic level under the null $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and asymptotic power under the sequence of contiguous alternatives $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{d}$ as defined in, for example, Hampel et al. (1986) and Ghosh et al. (2016b) among others. In particular, assuming that the contamination proportion tends to zero at the same rate in which the contiguous alternatives approaches to the null, here we consider the following contaminated joint distribution for the power stability calculation as

$$G^P_{n,\varepsilon,\boldsymbol{w}} = (1 - \tfrac{\varepsilon}{\sqrt{n}})G_{\boldsymbol{\beta}_n} + \tfrac{\varepsilon}{\sqrt{n}}\wedge_{\boldsymbol{w}}, \tag{4.2}$$

where $\boldsymbol{w}$ denote the contamination point $\boldsymbol{w} = (\boldsymbol{x}_t^T, y_t)^T$, and $G_{\boldsymbol{\beta}_n}$ denote the joint model distribution with true parameter value $\boldsymbol{\beta} = \boldsymbol{\beta}_n$. The contamination distribution to be considered for the level stability check can be obtained by substituting $\boldsymbol{d} = \boldsymbol{0}$ in (4.2), which yields

$$G^P_{n,\varepsilon,\boldsymbol{w}} = (1 - \tfrac{\varepsilon}{\sqrt{n}})G_{\boldsymbol{\beta}_0} + \tfrac{\varepsilon}{\sqrt{n}}\wedge_{\boldsymbol{w}}.$$

Then, the level and power influence functions are defined in terms of the following quantities

$$\alpha(\varepsilon, \boldsymbol{w}) = \lim_{n\to\infty} P_{G^L_{n,\varepsilon,\boldsymbol{w}}}(W_n > \chi^2_{r,\alpha}),$$

and

$$\pi(\boldsymbol{\beta}_n, \varepsilon, \boldsymbol{x}) = \lim_{n\to\infty} P_{G^P_{n,\varepsilon,\boldsymbol{w}}}(W_n > \chi^2_{r,\alpha}).$$

Definition 4.2. *The level influence function (LIF) and the power influence function (PIF) for the Wald-type test statistics W_n are defined respectively as*

$$\mathcal{LIF}(\boldsymbol{w}; W_n, G_{\boldsymbol{\beta}_0}) = \frac{\partial}{\partial\varepsilon}\alpha(\varepsilon, \boldsymbol{w})\Big|_{\varepsilon=0}, \quad \mathcal{PIF}(\boldsymbol{x}; W_n, G_{\boldsymbol{\beta}_0}) = \frac{\partial}{\partial\varepsilon}\pi(\boldsymbol{\beta}_n, \varepsilon, \boldsymbol{w})\Big|_{\varepsilon=0}.$$

See Ghosh et al. (2016b) for an extensive discussion on the interpretations of the level and power influence functions and their relations with the influence function of the test statistics in the context of a general Wald-type test.

Next, we will derive the forms of the LIF and PIF for our proposed tests in logistic regression model assuming the conditions required for the derivation of asymptotic distributions of the MDPDE hold.

Theorem 4.3. *Assume that the conditions of Theorem 6 hold and consider the contiguous alternatives $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{d}$ along with the contaminated model in (4.2). Then we have the following results:*

(i) The asymptotic distribution of the test statistics W_n under $G^P_{n,\varepsilon,\boldsymbol{w}}$ is non-central chi-square with r degrees of freedom and the non-centrality parameter

$$\delta = \widetilde{\boldsymbol{d}}^T_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0)\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0),$$

where $\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0) = \boldsymbol{d} + \varepsilon\mathcal{IF}(\boldsymbol{w}, \boldsymbol{T}_\lambda, G_{\boldsymbol{\beta}_0})$.

(ii) The asymptotic power under $G^P_{n,\varepsilon,\boldsymbol{w}}$ can be approximated as

$$\begin{aligned}\pi(\boldsymbol{\beta}_n,\varepsilon,\boldsymbol{w}) &\cong P\left(\chi^2_r(\delta)>\chi^2_{r,\alpha}\right)\\ &\cong \sum_{v=0}^{\infty} C_v\left(\boldsymbol{M}^T\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0),\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right)^{-1}\right)\\ &\quad\times P\left(\chi^2_{r+2v}>\chi^2_{r,\alpha}\right), \end{aligned} \tag{4.3}$$

where

$$C_v\left(\boldsymbol{t},\boldsymbol{A}\right)=\frac{\left(\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{t}\right)^v}{v!2^v}e^{-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{t}},$$

$\chi^2_p(\delta)$ denotes a non-central chi-square random variable with p degrees of freedom and δ as non-centrality parameter and $\chi^2_q=\chi^2_q(0)$ denotes a central chi-square random variable having degrees of freedom q.

Proof. Let us denote $\boldsymbol{\beta}^*_n=\boldsymbol{T}_\lambda(G^P_{n,\varepsilon,\boldsymbol{w}})$. Then, we get

$$\begin{aligned}W_n &= n(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{m})^T\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right)^{-1}(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{m})\\ &= n\left(\boldsymbol{M}^T\boldsymbol{\beta}^*_n-\boldsymbol{m}\right)^T\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right)^{-1}\left(\boldsymbol{M}^T\boldsymbol{\beta}^*_n-\boldsymbol{m}\right)\\ &\quad+n(\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*_n)^T\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T(\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*_n)\\ &\quad+n(\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*_n)^T\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right)^{-1}(\boldsymbol{M}^T\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{m})\\ &= S_{1,n}+S_{2,n}+S_{3,n}. \end{aligned} \tag{4.4}$$

Next, one can show that

$$\begin{aligned}\sqrt{n}(\boldsymbol{\beta}^*_n-\boldsymbol{\beta}_0) &= \boldsymbol{d}+\varepsilon\mathcal{IF}\left(\boldsymbol{w},\boldsymbol{T}_\lambda,G_{\boldsymbol{\beta}_0}\right)+o_p(\mathbf{1}_p)\\ &= \widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\theta}_0)+o_p(\mathbf{1}_p). \end{aligned} \tag{4.5}$$

Thus, we get

$$\sqrt{n}(\boldsymbol{M}^T\boldsymbol{\beta}^*_n-\boldsymbol{m})=\boldsymbol{M}^T\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\theta}_0)+o_p(\mathbf{1}_p). \tag{4.6}$$

Further, under $G^P_{n,\varepsilon,\boldsymbol{w}}$, the asymptotic distribution of MDPDE yields

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*_n)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\right). \tag{4.7}$$

Thus, we get

$$S_{3,n}\xrightarrow[n\to\infty]{\mathcal{L}}\chi^2_r.$$

Combining (4.4), (4.6) and (4.7), we get

$$W_n=\boldsymbol{Z}_n^T\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right)^{-1}\boldsymbol{Z}_n+o_p(1),$$

where

$$\boldsymbol{Z}_n=\sqrt{n}\boldsymbol{M}^T(\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*_n)+\boldsymbol{M}^T\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\theta}_0).$$

By (4.7),

$$\boldsymbol{Z}_n\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(\boldsymbol{M}^T\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\theta}_0),\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda\left(\boldsymbol{\beta}_0\right)\boldsymbol{M}\right),$$

and hence we get that

$$W_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_r(\delta),$$

where δ is as defined in Part (i) of the theorem.

Part (ii) of the theorem follows from Part (i) using the infinite series expansion of a non-central chi-square distribution function in terms of that of the central chi-square variables:

$$\begin{aligned}\pi(\boldsymbol{\beta}_n, \varepsilon, \boldsymbol{w}) &= \lim_{n\to\infty} P_{G^P_{n,\varepsilon,\boldsymbol{w}}}(W_n > \chi^2_{r,\alpha}) \cong P(\chi^2_{r,\delta} > \chi^2_{r,\alpha}) \\ &= \sum_{v=0}^{\infty} C_v \left(\boldsymbol{M}^T \widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0), \left(\boldsymbol{M}^T \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\, \boldsymbol{M}\right)^{-1}\right) P\left(\chi^2_{r+2v} > \chi^2_{r,\alpha}\right).\end{aligned}$$

□

Corollary 4.1. *Putting $\varepsilon = 0$ in Theorem 4.3, we get the asymptotic power of the proposed Wald-type tests under the contiguous alternative hypotheses $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{d}$ as*

$$\pi(\boldsymbol{\beta}_n) = \pi(\boldsymbol{\beta}_n, 0, \boldsymbol{w}) \cong \sum_{v=0}^{\infty} C_v \left(\boldsymbol{M}^T \boldsymbol{d}, \left(\boldsymbol{M}^T \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\, \boldsymbol{M}\right)^{-1}\right) P\left(\chi^2_{r+2v} > \chi^2_{r,\alpha}\right).$$

This is identical with the results obtained earlier in Theorem 10 independently.

Corollary 4.2. *Putting $\boldsymbol{d} = \boldsymbol{0}$ in Theorem 4.3, we get the asymptotic distribution of W_n under $G^L_{n,\varepsilon,\boldsymbol{w}}$ as the non-central chi-square distribution having r degrees of freedom and non-centrality parameter*

$$\varepsilon^2 \mathcal{IF}(\boldsymbol{w}; \boldsymbol{T}_\lambda, G_{\boldsymbol{\beta}_0})^T \boldsymbol{M} \left(\boldsymbol{M}^T \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\, \boldsymbol{M}\right)^{-1} \boldsymbol{M}^T \mathcal{IF}(\boldsymbol{w}; \boldsymbol{T}_\lambda, G_{\boldsymbol{\beta}_0}).$$

Then, the asymptotic level under contiguous contamination is given by

$$\begin{aligned}&\alpha(\varepsilon, \boldsymbol{w}) = \pi(\boldsymbol{\beta}_0, \varepsilon, \boldsymbol{w}) \\ &\cong \sum_{v=0}^{\infty} C_v \left(\varepsilon \boldsymbol{M}^T \mathcal{IF}(\boldsymbol{w}; \boldsymbol{T}_\lambda, G_{\boldsymbol{\beta}_0}), \left(\boldsymbol{M}^T \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\, \boldsymbol{M}\right)^{-1}\right) P\left(\chi^2_{r+2v} > \chi^2_{r,\alpha}\right).\end{aligned}$$

In particular, as $\varepsilon \to 0, \boldsymbol{\beta}^*_n \to \boldsymbol{\beta}_0$ and the non-centrality parameter of the above asymptotic distribution tends to zero leading to the null distribution of W_n.

Now we can easily obtain the the power and level influence functions of the Wald-type test statistics from Theorem 4.3 and Corollary 4.2 and these have been presented in the following theorem.

Theorem 4.4. *Under the assumptions of Theorem 4.3, the power and level influence functions of the proposed Wald-type test statistic W_n is given by*

$$\mathcal{PIF}(\boldsymbol{w}, W_n, G_{\boldsymbol{\beta}_0}) \cong K^*_r\left(\boldsymbol{s}^T(\boldsymbol{\beta}_0)\, \mathbf{d}\right) \boldsymbol{s}^T(\boldsymbol{\beta}_0) \mathcal{IF}(\boldsymbol{w}, \boldsymbol{T}_\lambda, G_{\boldsymbol{\beta}_0}), \tag{4.8}$$

with $\boldsymbol{s}^T(\boldsymbol{\beta}_0) = \boldsymbol{d}^T\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T$ *and*

$$K_r^*(s) = e^{-\frac{s}{2}}\sum_{v=0}^{\infty}\frac{s^{v-1}}{v!2^v}(2v-s)\,P\left(\chi^2_{r+2v} > \chi^2_{r,\alpha}\right),$$

and

$$\mathcal{LIF}(\boldsymbol{w}, W_n, G_{\boldsymbol{\beta}_0}) = 0.$$

Further, the derivative of $\alpha(\varepsilon, \boldsymbol{w})$ *of any order with respect to* ε *will be zero at* $\varepsilon = 0$*, implying that the level influence function of any order will be zero.*

Proof. We start with the expression of $\pi(\boldsymbol{\beta}_n, \varepsilon, \boldsymbol{w})$ from Theorem 4.3. Clearly, by definition of PIF and using the chain rule of derivatives, we get

$$\begin{aligned}
&\mathcal{PIF}(\boldsymbol{w}, W_n, G_{\boldsymbol{\beta}_0}) = \frac{\partial}{\partial\varepsilon}\,\pi(\boldsymbol{\beta}_n, \varepsilon, \boldsymbol{w})\big|_{\varepsilon=0}\\
&\cong \sum_{v=0}^{\infty}\frac{\partial}{\partial\varepsilon}\,C_v\left(\boldsymbol{M}^T\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0), \left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1}\right)\Big|_{\varepsilon=0} P\left(\chi^2_{r+2v} > \chi^2_{r,\alpha}\right)\\
&\cong \sum_{v=0}^{\infty}\frac{\partial}{\partial\boldsymbol{t}^T}\,C_v\left(\boldsymbol{M}^T\boldsymbol{t}, \left(\boldsymbol{M}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{M}\right)^{-1}\right)\Big|_{\boldsymbol{t}=\widetilde{\boldsymbol{d}}_{0,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0)}\\
&\quad\times\frac{\partial}{\partial\varepsilon}\,\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0)\Big|_{\varepsilon=0} P\left(\chi^2_{r+2v} > \chi^2_{r,\alpha}\right).
\end{aligned}$$

Now $\widetilde{\boldsymbol{d}}_{0,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0) = \boldsymbol{d}$ and standard differentiations give

$$\frac{\partial}{\partial\varepsilon}\widetilde{\boldsymbol{d}}_{\varepsilon,\boldsymbol{w},\lambda}(\boldsymbol{\beta}_0) = \mathcal{IF}(\boldsymbol{w}, \boldsymbol{T}_\lambda, G_{\boldsymbol{\beta}_0}),$$

and

$$\frac{\partial}{\partial\boldsymbol{t}}C_v(\boldsymbol{t}, \boldsymbol{A}) = \frac{\left(\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{t}\right)^{v-1}}{v!2^v}\left(2v - \boldsymbol{t}^T\boldsymbol{A}\boldsymbol{t}\right)\boldsymbol{A}\boldsymbol{t}e^{-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{t}}.$$

Combining above results and simplifying, we get the required expression of PIF as presented in the theorem. □

It is clear from the above theorem that, the asymptotic level of the proposed Wald-type test statistic will be unaffected by a contiguous contamination for any values of the tuning parameter λ, whereas the power influence function will be bounded whenever the influence function of the MDPDE is bounded (which happens for all $\lambda > 0$). Thus, the robustness of the power of the proposed tests again turns out to be directly dependent on the robustness of the MDPDE $\boldsymbol{\beta}_\lambda$ used in constructing the test. In particular, the asymptotic contiguous power of the classical MLE based Wald-type test (at $\lambda = 0$) will be non-robust whereas that for the Wald-type tests with $\lambda > 0$ will be robust under contiguous contamination and this robustness increases as λ increases further.

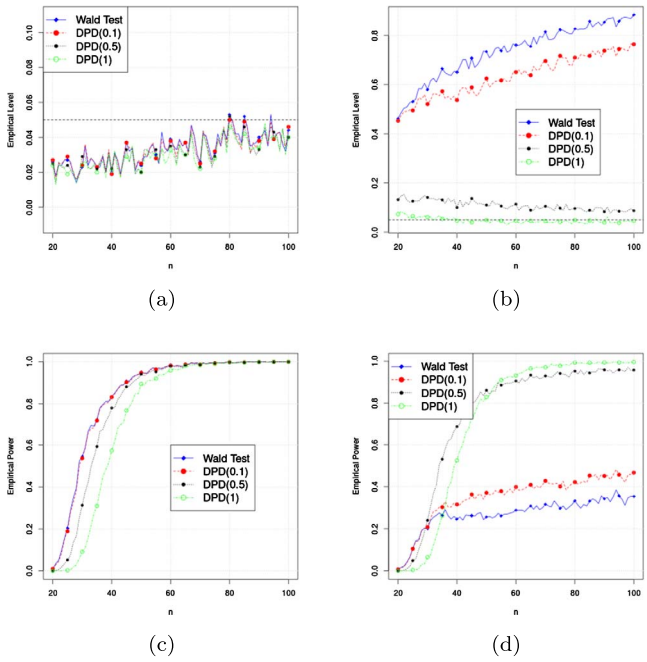


FIG 4. *(a) Simulated levels of different tests for pure data; (b) simulated levels of different tests for contaminated data; (c) simulated powers of different tests for pure data; (d) simulated powers of different tests for contaminated data.*

5. Simulation study

In this section we have empirically demonstrated some of the strong robustness properties of the density power divergence tests for the logistic regression model. We considered two explanatory variables x_1 and x_2 in this study, so $k = 2$. These two variables are distributed according to a standard normal distribution $\mathcal{N}(\mathbf{0},\boldsymbol{I}_{2\times 2})$. The response variables Y_i are generated following the logit model as given in (1.1). The true value of the parameter is taken as $\boldsymbol{\beta}_0 = (0, 1, 1)^T$. We considered the null hypothesis $H_0 : (\beta_1, \beta_2)^T = (1, 1)^T$. It can be written in the form of the general hypothesis given in (1.3), where $\boldsymbol{m} = (1, 1)^T$ and

$$\boldsymbol{M} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Our interest was in studying the observed level (measured as the proportion of test statistics exceeding the corresponding chi-square critical value in a large number – here 1000 – of replications) of the test under the correct null hypothesis. The result is given in Figure 4(a) where the sample size n varies from 20 to 100. We have used several Wald-type test statistics, corresponding to different minimum density power divergence estimators. We have used, $\lambda = 0,\ 0.1,\ 0.5$

and 1, in this particular study. As it is previously mentioned, $\lambda = 0$ is the classical Wald test for the logistic regression model. The horizontal lines in the figure represents the nominal level of 0.05. It may be noticed that all the tests are slightly conservative for small sample sizes and lead to somewhat deflated observed levels. In particular, the Wald-type tests with higher values of λ are relatively more conservative. However, this discrepancy decreases rapidly as sample size increases.

To evaluate the stability of the level of the tests under contamination, we repeated the tests for the same null hypothesis by adding 3% outliers in the data. For the outlying observations we first introduced the leverage points where x_1 and x_2 are generated from $\mathcal{N}(\boldsymbol{\mu}_c, \sigma \boldsymbol{I}_{2\times 2})$ with $\boldsymbol{\mu}_c = (5,5)^T$ and $\sigma = 0.01$. Then the values of the response variable corresponding to those leverage points were altered to produce vertical outliers ($y_t = 1$ was converted to $y_t = 0$). Figure 4(b) shows that the levels of the classical Wald test as well as DPD(0.1) test break down, whereas Wald-type test statistics for $\lambda = 0.5$ and $\lambda = 1$ present highly stable levels.

To investigate the power of the tests we changed the null hypothesis to H_0^* : $(\beta_1, \beta_2)^T = (0,0)^T$, and kept the data generating distributions as before, as well as the true value of the parameter as $\boldsymbol{\beta}_0 = (0,1,1)^T$. In terms of the null hypothesis in (1.3) the value of $\boldsymbol{m}$ is changed to $(0,0)^T$ whereas $\boldsymbol{M}$ remained unchanged from the previous experiment. The empirical power functions are calculated in the same manner as the levels of the tests, and plotted in Figure 4(c). The Wald test is the most powerful under pure data. The power of the Wald-type test statistic for $\lambda = 0.1$ almost coincides with the classical Wald test in this case. The performances of the Wald-type test statistics for $\lambda = 0.5$ and $\lambda = 1$ are relatively poor, however, as the sample size increases to 60 and beyond, the powers are practically identical.

Finally, we calculated the power functions under contamination for the above hypothesis under the same setup as that of the level contamination. The observed powers of that the tests are given in Figure 4(d). The Wald-type test statistics for $\lambda = 0.5$ and $\lambda = 1$ show stable powers under contamination, but the classical Wald test and the Wald-type test for $\lambda = 0.1$ exhibit a drastic loss in power. In very small sample sizes the classical Wald test and the Wald-type test for $\lambda = 0.1$ have slightly higher power than the other tests, but this advantage quickly disppears with increasing sample size. On the whole, the proposed Wald-type test statistics corresponding to moderately large λ appear to be quite competitive to the classical Wald test for pure normal data, but they are far better in terms of robustness properties under contaminated data.

6. Real data examples

In this section we will explore the performance of the proposed Wald-type tests in logistic regression models by applying it on different interesting real data sets. The estimators are computed by minimizing the corresponding density power divergence through the software `R`, and the minimization is performed using "optim" function.

6.1. Students data

As an interesting data example leading to the logistic regression model, we consider the students data set from Muñoz-Garcia et al. (2006). The data set consists of 576 students of the University of Seville. The response variable is the students aim to graduate after three years. The explanatory variables are gender ($x_{i1} = 0$ if male; $x_{i1} = 1$ if female), entrance examination (EE) in University ($x_{i2} = 1$ if the first time; $x_{i2} = 0$ otherwise) and sum of marks (x_{i3}) obtained for the courses of first term. There were 61 distinct cases (i.e. $n = 61$) in this study. We assume that the response variable follows a binomial logistic regression model as mentioned in Remark 2.1. We are interested to test the null hypothesis that the gender of student does not play any role on their aim. So the null hypothesis is given by $H_0 : \beta_1 = 0$. Figure 5 shows p-values of Wald-type tests for different values of λ. Muñoz-Garcia et al. (2006) mentioned that the 32nd observation is the most influential point as it has a large residual and a high leverage value. If we use the classical Wald test or Wald-type tests with small λ under the full data, the null hypothesis is rejected at 10% level of significance. But this result is clearly a false positive as the outlier deleted p-values for all λ are close to 0.35. On the other hand, Wald-type tests with large λ give robust p-values in both situations.

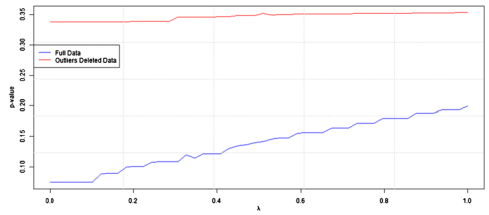


FIG 5. *P-values of Wald-type tests for testing* $H_0 : \beta_1 = 0$ *in Students data.*

6.2. Lymphatic cancer data

Brown (1980), Martín and Pardo (2009) and Zelterman (2005, Section 3.3) studied the data that focused on the evidence of lymphatic cancer in prostate cancer patients for predicting lymph nodal involvement of cancer. There were five covariates (three dichotomous and two continuous): the X-ray finding ($x_{i1} = 1$ if present; $x_{i1} = 0$ if absent), size of the tumor by palpation ($x_{i2} = 1$ if serious; $x_{i2} = 0$ if not serious), pathology grade by biopsy ($x_{i3} = 1$ if serious; $x_{i3} = 0$ if not serious), the age of the patient at the time of diagnosis (x_{i4}) and serum acid phosphatase level (x_{i5}). The diagnostics was associated with 53 individuals. An ordinary logistic model is assumed here. We are interested in testing the significance of the size of the tumor on the response variable, so the null hypothesis is taken as $H_0 : \beta_2 = 0$. The p-values of Wald-type tests for different values of λ are given in Figure 6. Martín and Pardo (2009) noticed that the 24th observation is an influential point. The p-value of the classical Wald test

under the full data is 0.0430, but if the outlier is deleted it becomes 0.0668. So if we consider a test at 5% level of significance, the decision of the test changes when we delete just one outlying observation. However, Wald-type tests with high values of λ always produce high p-values.

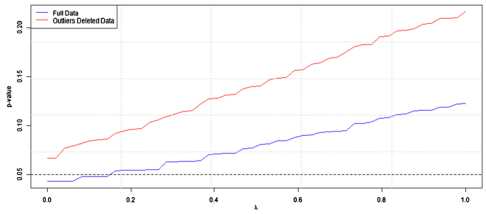


FIG 6. *P-values of Wald-type tests for testing* $H_0 : \beta_2 = 0$ *in Lymphatic Cancer data.*

6.3. Vasoconstriction data

Finney (1947), Pregibon (1981) and Martín and Pardo (2009) studied the data where the interest is in the occurrence of vasoconstriction in the skin of the finger. The covariates of the study were the logarithm of volume (x_{i1}) and the logarithm of rate (x_{i2}) of inspired air measured in liters. Pregibon (1981) has shown that two observations, the 4th and 18th, are not fitted well by the logistic model as they have large residuals. However, it can be checked easily that these observations are only outliers in the y-space and are not leverage points. Here we want to test that there is no effect of the covariates, so the null hypothesis is given by $H_0 : \beta_1 = \beta_2 = 0$. The p-value of the classical Wald test under the full data is 0.0194, and in the outlier deleted data it becomes 0.0371. But, Figure 7 shows that Wald-type tests with large λ produce large p-values.

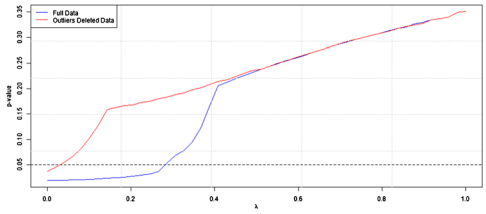


FIG 7. *P-values of Wald-type tests for testing* $H_0 : \beta_1 = \beta_2 = 0$ *in Vasoconstriction data.*

6.4. Leukemia data

This data set consists of 33 cases on the survival of individuals diagnosed with leukemia. The explanatory variables are white blood cell count (x_{i1}) and another variable which indicates the presence or absence of a certain morphological characteristic in the white cells ($x_{i2} = 1$ if present; $x_{i2} = 0$ if absent). This data set was also studied by Cook and Weisberg (1982), Johnson (1985) and Martín

and Pardo (2009). They defined a success to be patient survival in excess of 52 weeks. We are interested to test the significance of two covariates, i.e. the null hypothesis is $H_0 : \beta_1 = \beta_2 = 0$. The plot of the p-values of Wald-type tests for different values of λ is given in Figure 8. Martín and Pardo (2009) noticed that the 15th observation is an influential point. The p-value of the classical Wald test under the full data is 0.0226, but if the outlier is deleted it becomes 0.0683. Thus, at 5% level of significance, the decision of the test depends on only one outlying observation. In this case also Wald-type tests with high values of λ always produce high p-values.

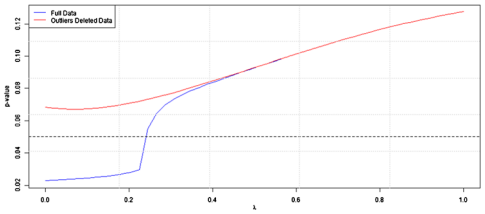


FIG 8. *P-values of Wald-type tests for testing* $H_0 : \beta_1 = \beta_2 = 0$ *in Leukemia data.*

7. On the choice of tuning parameter λ

In this paper, we have proposed a robust family of Wald-type test statistics for testing general linear hypothesis under the logistic regression model, which depend crucially on a tuning parameter λ involved in its definition. We have seen from all the theoretical results and numerical illustrations throughout the paper that the power for contiguous alternative hypotheses for the proposed Wald-type tests decrease slightly with increasing λ under pure data with no contamination but, on the other hand, in presence of contamination in data the stability of both power and level increases drastically with increasing λ. In particular, it can be noted from Figure 4 that the loss in power is not very significant even for $\lambda \approx 0.5$ under moderate sample size and this loss becomes almost zero for larger sample sizes; however levels of the tests are highly stable in presence of contamination for any sample size with $\lambda \approx 0.5$. Further, from the real data examples (Figures 5–8), we can also see that the p-values and the resulting inferences are highly robust for $\lambda \geq 0.4$. All this empirical evidences suggest the use of $\lambda \approx 0.5$ as an ad-hoc choice of tuning parameter while applying the proposed method in any practical problem and is expected to produce a fair enough trade-off between the power under pure data and robustness under contamination.

Although this ad-hoc choice of λ works quite well in practice, many practitioner may believe that the level of contamination is different for each practical data set and hence we should have different trade-off for each of them. This can be done through an appropriate algorithm to obtain an data-driven choice of λ separately for each sample, which provides a trade-off between true power and robustness against outliers for the test based on only the given data at hand. To develop such a method for our proposed Wald-type test statistics, we note

that their performances are directly dependent on that of the MDPDE having the same tuning parameter λ. The power of our proposed Wald-type tests at contiguous alternatives, as obtained in Theorem 10, increases whenever the non-centrality parameter $\delta = \boldsymbol{d}^T \boldsymbol{\Sigma}_\lambda (\boldsymbol{\beta}_0) \boldsymbol{d}$ of the associated chi-square distribution decreases, i.e., whenever the asymptotic variance $\boldsymbol{\Sigma}_\lambda (\boldsymbol{\beta}_0)$ of the MDPDE used decreases and its asymptotic efficiency increases. Hence, as λ increases, both the asymptotic power of the Wald-type test at contiguous alternatives and the asymptotic efficiency of the MDPDE decreases slightly. On the other hand the influence function of the Wald-type test statistics as well as its power influence functions are a direct function of the influence function of the MDPDE with the same tuning parameter. Therefore the robustness of both the test and the MDPDE are equivalently dependent of their respective tuning parameter λ; in particular, their robustness increases significantly with increasing λ. Therefore, the problem of a suitable data-driven selection of the tuning parameter λ for the proposed Wald-type test statistics through proper trade-off between its power under true data and robustness can be equivalently solved by obtaining a data-driven tuning parameter with proper trade-off of asymptotic efficiency and robustness of the MDPDE used in constructing the test statistics.

There are a few existing approaches of selection of data-driven tuning parameter of the general MDPDE under i.i.d. setup; among them the popular one is the method of Warwick and Jones (2005) who proposed the minimization of an estimator of the MSE of the MDPDE to get optimum λ. The approach has been recently studied in many contexts with suitable extensions (Ghosh and Basu, 2013; 2015; Ghosh et al., 2016a, 2016c) and shown to provide satisfactory performances in selecting proper tuning parameter for any given data set. Here, we will use their approach to propose a data-driven selection of the tuning parameter λ of the MDPDE and hence for the proposed Wald-type tests under the present logistic regression model. Following Warwick and Jones (2005), we need to minimize an estimate of the MSE of the MDPDE $\widehat{\boldsymbol{\beta}}_\lambda$ as an function of the tuning parameter λ given by

$$MSE(\lambda) = (\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}^*)^T (\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}^*) + \frac{1}{n} \text{Trace} \left(\boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_\lambda) \boldsymbol{K}_\lambda(\boldsymbol{\beta}_\lambda) \boldsymbol{J}_\lambda^{-1}(\boldsymbol{\beta}_\lambda) \right), \tag{7.1}$$

where $\boldsymbol{\beta}^*$ is the true value of the target parameter and $\boldsymbol{\beta}_\lambda$ is the best fitting parameter that minimizes the DPD measure (with tuning parameter λ) between the true and the (model density. Note that, although we have considered the model to be correct in the previous parts of the paper, the above construction gives us more flexibility to work with true densities outside the model family also. In particular, the first term in (7.1) indicates the model misspecification bias and becomes zero whenever the true density belongs to the assumed model family. On the other hand, the second term in (7.1) simply gives the variance of the MDPDE. We need to get an estimate of this MSE based on the given data without assuming that the model is true and then minimize this suitable oven $\lambda \in [0, 1]$ (may be through a grid search) to get the optimum λ for the data at hand.

In order to estimate the MSE in (7.1), let us first consider the (second) variance term. We have already provided estimator of $\boldsymbol{J}_\lambda(\boldsymbol{\beta}_\lambda)$ and $\boldsymbol{K}_\lambda(\boldsymbol{\beta}_\lambda)$ in Section 2 but assuming that the model is true. One can easily obtain their model free estimators also in a similar way, which could be given by

$$\widehat{\boldsymbol{J}}_\lambda^* = \frac{1}{n}\sum_{i=1}^{n}\left[L_1(\boldsymbol{x}_i, y_i, \widehat{\boldsymbol{\beta}}_\lambda) + L_2(\boldsymbol{x}_i, y_i, \widehat{\boldsymbol{\beta}}_\lambda)\right],$$

$$\widehat{\boldsymbol{K}}_\lambda^* = \frac{1}{n}\sum_{i=1}^{n}\frac{(e^{\lambda \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda} + e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda})^2}{(1+e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda})^{2(\lambda+2)}}\left(e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda} - y_i(1+e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda})\right)^2 \boldsymbol{x}_i\boldsymbol{x}_i^T.$$

where L_1 and L_2 are defined in Section 2. Next, in order to estimate the (first) bias term in (7.1), we can estimate $\boldsymbol{\beta}_\lambda$ by the MDPDE $\widehat{\boldsymbol{\beta}}_\lambda$ but there is no obvious choice for $\boldsymbol{\beta}^*$. Warwick and Jones (2005) suggested to use suitable pilot estimator $\boldsymbol{\beta}^P$ in place of $\boldsymbol{\beta}^*$ and use the following estimate of the MSE:

$$\widehat{MSE}(\lambda) = \left(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^P\right)^T\left(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^P\right) + \frac{1}{n}\text{Trace}\left(\widehat{\boldsymbol{J}}_\lambda^{*-1}\widehat{\boldsymbol{K}}_\lambda^*\widehat{\boldsymbol{J}}_\lambda^{*-1}\right).$$

Note that, this selection procedure clearly depends on the pilot estimator used. When we take the pilot estimator $\boldsymbol{\beta}^P = \widehat{\boldsymbol{\beta}}_\lambda$, it corresponds to the assumption of no model bias and the approach coincides with that of Hong and Kim (2001); this is clearly more restrictive and we lose generality of the procedure against outliers due to model misspecification. Alternatively, Warwick and Jones (2005) suggested, through an extensive simulation study, that the choice $\boldsymbol{\beta}^P = \widehat{\boldsymbol{\beta}}_1$ works well enough for the case of MDPDE under i.i.d. data. Later Ghosh and Basu (2015) empirically concluded, while extending to the non-homogeneous setup, the choice $\boldsymbol{\beta}^P = \widehat{\boldsymbol{\beta}}_{0.5}$ often works better. Here, we will first empirically examine a good choice of the pilot estimator for the present case of random design logistic regression and illustrate that this method works in practice for choosing a data-driven choice of tuning parameter λ.

Let us reconsider the simulation study discussed in Section 5, but now we perform the selection of λ following the above proposal for each iteration with different possible pilot estimators. Figure 9 gives the simulated level and power in the same setup of Figure 4. The average optimum values of λ for the pure data as well as the contaminated data are plotted in Figure 10. Pilot(λ) in these plots refers to the Wald-type test statistic where MDPDE with the tuning parameter λ is used as a pilot estimator. Figure 9 (a) and (c) show that in the pure data there is no significant effect of the pilot estimator to the level and power of the tests. In fact, Figure 10 (a) shows that the optimum λ turns out to be small (less than 0.25) in case of the pure data. This result is consistent with the result in Figure 4 (a) and (c) as we noticed almost no difference in the level or power of the tests with small values of λ. However, in the contaminated data the tuning parameter plays a vital role in the robustness of the test. This is also true for the pilot estimator. Figure 10 (b) shows that the optimum value of λ is still small if a pilot estimator with small λ is chosen. As a result, the level of the test breaks down and the power of the test is not sufficiently high.

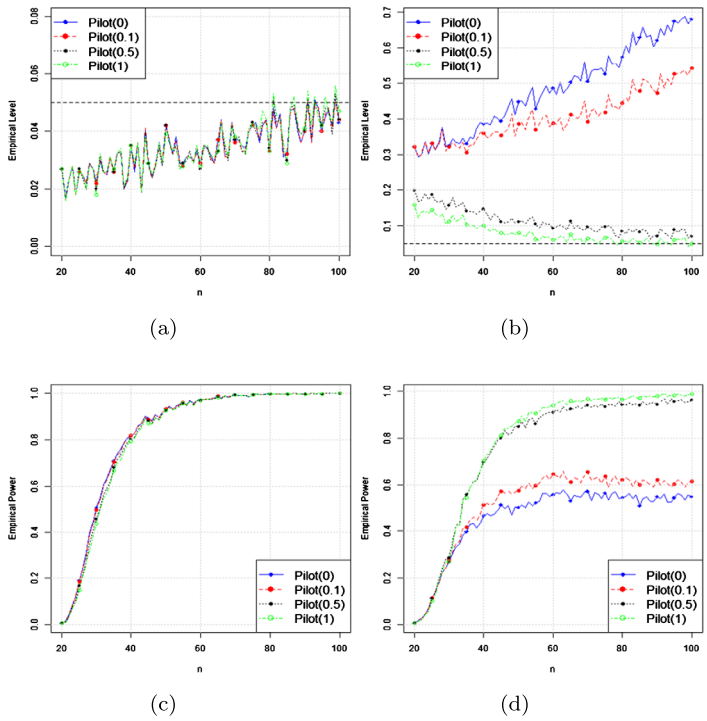


FIG 9. *(a) Simulated level and power using the optimum λ for different values of the pilot estimator: (a) simulated levels for pure data, (b) simulated levels for contaminated data; (c) simulated powers for pure data; (d) simulated powers contaminated data.*

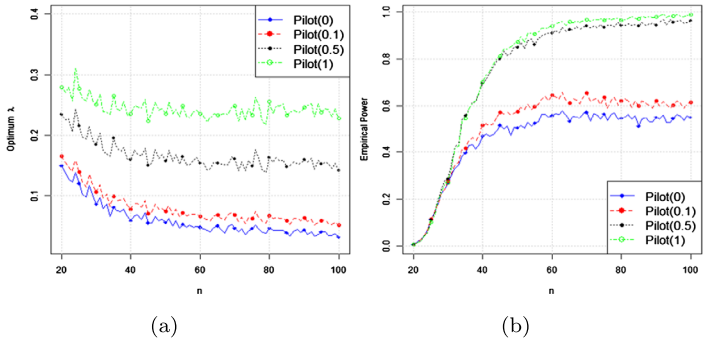


FIG 10. *The average optimum values of λ for different values of the pilot estimator for (a) the pure data and (b) the contaminated data.*

On the other hand, a pilot estimator with large λ produces a large optimum λ (more than 0.5), so the corresponding test gives a stable level and high power, see Figure 9 (b) and (d). The simulation results indicate that the performance

Students Data						
Pilot λ	0	0.1	0.25	0.5	0.75	1
Optimum λ	1	1	1	1	1	1
Optimum MSE	10.21	10.21	10.11	10.03	9.97	9.98
Lymphatic Cancer Data						
Pilot λ	0	0.1	0.25	0.5	0.75	1
Optimum λ	0	0	0	0	0	0
Optimum MSE	13.57	13.57	13.59	13.65	13.71	13.75
Vasoconstriction Data						
Pilot λ	0	0.1	0.25	0.5	0.75	1
Optimum λ	1	1	1	1	1	1
Optimum MSE	50.76	45.41	24.51	15.04	15.05	15.04
Leukemia Data						
Pilot λ	0	0.1	0.25	0.5	0.75	1
Optimum λ	0.47	0.47	0.47	0.47	0.47	0.47
Optimum MSE	5.14	4.98	3.01	2.98	2.98	2.98

TABLE 1

The optimum values of λ and MSE using different pilot estimators for different data sets.

Data Set	Optimum λ	p-value^1	p-value^2
Students Data	1	0.1992	0.3539
Lymphatic Cancer Data	0	0.0430	0.0668
Vasoconstriction Data	1	0.3506	0.3506
Leukemia Data	0.47	0.0900	0.0903

TABLE 2

The optimum values of λ and the corresponding p-values in full data (p-value^1) and in outliers deleted data (p-value^2).

of the tests with pilot estimators $\lambda = 0.5$ and 1 both give sufficient robustness properties, moreover, $\lambda = 1$ gives slightly better results in this simulation setup. The similar scenario is observed in case of the DPD test with a fixed value of λ.

Now, we apply the proposed method of optimal selection of tuning parameter λ to all our real data examples of Section 6. We use several pilot estimators, but finally, they produced almost same optimal λ; see in Table 1 the detailed results. We have also computed the p-values corresponding to these optimum value of λ in the full data and in outliers deleted data for each examples, which are reported in Table 2. Note that, the p-values do not change significantly in the presence of outliers when the tuning parameter λ is chosen optimally for each example following the proposed algorithm. The interpretation of the result remains the same as we discussed in the previous section; however, as it is based on the optimum choice of λ, it eliminates the subjective choice of the tuning parameter for the DPD tests.

8. Concluding remarks

Logistic regression for binary outcomes is one of the most popular and successful tools in the statisticians toolbox. It is frequently used by applied scientists of many disciplines to solve problems of real interest in their domain of application. However, in the present age of big data, the need for protection against

data contamination and other modeling errors is paramount, and, wherever possible, strong robustness qualities should be a default requirement for statistical methods used in practice. In this paper we have presented one such class of inference procedures. We have provided a thorough theoretical evaluation of the proposed class of tests for testing the linear hypothesis in the logistic regression model highlighting their robustness advantages. We have also produced substantial numerical evidence, including simulation results and a large number of real problems, to demonstrate how these theoretical advantages translate in practice to real gains. On the whole, we feel that the proposed tests will turn out to be an useful set of tools with significant practical application.

References

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. MR1665873

Basu, A., Shioya, H. and Park, C. (2011). *The minimum distance approach. Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton. MR2830561

Basu, A., Mandal, A., Martín, N. and Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics*, **65**, 319–348. MR3011625

Basu, A., Mandal, A., Martín, N. and Pardo, L. (2015). Robust tests for the equality of two normal means based on the density power divergence. *Metrika,* **78**, 611–634. MR3355464

Basu, A., Mandal, A., Martín, N., Pardo, L. (2016). Generalized Wald-type tests based on minimum density power divergence estimators. *Statistics,* **50**, 1–26. MR3435166

Bianco, A. M. and Martinez, E. (2009). Robust testing in the logistic regression model. *Computational Statistics and Data Analysis*, **53**, 4095–4105. MR2744307

Bianco, A. M. and Yohai, V. J. (1996). Robust Estimation in the Logistic Regression Model,in Robust Statistics, Data Analysis and Computer Intensive Methods, 17–34; Lecture Notes in Statistics 109, Springer Verlag, Ed. H. Rieder. New York MR1491394

Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika*, **92**, 724–731. MR2202658

Bondell, H. D. (2008). A characteristic function approach to the biased sampling model, with application to robust logistic regression. *Journal of Statistical Planning and Inference,* **138**, 742–755. MR2382886

Brown, B. W. (1980). Prediction analysis for binary data, in *Biostatistics Casebook*, R. G. Miller, B. Efron, B. W. Brown and L. E. Moses, eds., John Wiley and Sons, New York, pp. 3–18.

Carroll, R. J. and Pederson, S. (1993). On Robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B,* **55**, 669–706. MR1223937

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society: Series B,* **50**, 225–265.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, **44**, 273–295. MR2020151

Christmann, A. (1994). Least Median of Weighted Squares in Logistic Regression with Large Strata. *Biometrika,* **81**, 413–417. MR1294903

Christmann, A. and Rousseeuw, P. J. (2001). Measuring overlap in binary regression, *Comp. Statistics & Data Analysis*, **37**, 65–75. MR1862480

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman & Hall, London. MR0675263

Feigl, P. and Zelen, M. (1965). Estimation of exponential probabilities with concomitant information. *Biometrics*, **21**, 826–838.

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, **34**, 320–334.

Ghosh, A. and Basu, A. (2013). Robust Estimation for Independent but Non-Homogeneous Observations using Density Power Divergence with application to Linear Regression. *Electronic Journal of Statistics*, **7**, 2420–2456. MR3117102

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 Statistics*, **42**(9), 2056–2072. MR3371040

Ghosh, A., Basu, A. and Pardo, L. (2015). On the robustness of a divergence based test of simple statistical hypotheses, *Journal of Statistical Planning and Inference,* **161**, 91–108. MR3316553

Ghosh, A., Harris, I. R., Maji, A., Basu, A. and Pardo, L. (2016a). A Generalized Divergence for Statistical Inference. *Bernoulli*, **23**(4A), 2746–2783. MR2732941

Ghosh, A., Mandal, A., Martín, N. and Pardo, L. (2016b). Influence Analysis of Robust Wald-type Tests. *Journal of Multivariate Analysis,* **147**, 102–126. MR3484172

Ghosh, A., Martin, N., Basu, A. and Pardo, L. (2016c). A New Class of Robust Two-Sample Wald-Type Tests. *eprint arXiv:1702.04552.*

A. Ghosh and A. Basu (2016d). Robust Estimation in Generalized Linear Models: The Density Power Divergence Approach. *TEST*, **25**(2), 269–290.

Greene, W. H. (2003). *Econometric Analysis.* Upper Saddle *River*: Prentice Hall Inc.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A., (1986). *Robust statistics: The approach based on influence functions.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. MR0829458

Hobza, T., Pardo, L. and I. Vajda (2008). Robust median estimator in logistic regression. *Journal of Statistical Planning and Inference,* **138**, 3822–3840. MR2455970

Hobza, T., Martín, N. and Pardo, L. (2017). A Wald-type test statistic based

on robust modified median estimator in logistic regression models. *Journal of Statistical Computation and Simulation,* **87**, 2309–2333.

Hong, C. and Y. Kim (2001). Automatic selection of the tuning parameter in the minimum density power divergence estimation. *Journal of the Korean Statistical Association*, **30**, 453–465. MR1895987

Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrics,* **72**, 59–65.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics. Theory and Methods*. Wiley Series in Probability and Statistics. MR2238141

Martín, N. and Pardo, L. (2009). On the asymptotic distribution of Cook's distance in logistic regression models. *Journal of Applied Statistics*, **36**, 1119–1146. MR2744128

Muñoz-Garcia, J., Muñoz-Pichardo, J. M. and Pardo, L. (2006). Cressie and Read power-divergences as influence measures for logistic regression models. *Comput. Statist. Data Anal.*, **50**, 3199–3221. MR2239664

Morgenthaler, S. (1992), Least-absolute-deviations fits for generalized linear models. *Biometrika,* **79**, 747–754.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705–724. MR0619277

Pregibon, D. (1982), Resistant lits for some commonly used logistic models with medical applications, *Biometrics*, **38,** 485–498.

Rousseeuw, P. J. and Christmann, A. (2003), Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis,* **43**, 315–332. MR1996815

Rousseeuw, P. J. and Ronchetti, E. (1979) The influence curve for tests. *Research Report* 21, Fachgruppe fur Statistik, ETH Zurich.

Toma, A. and Broniatowski, M. (2011). Dual divergence estimators and tests: Robustness results. *Journal of Multivariate Analysis*, **102**, 20–36. MR2729417

Victoria-Feser, M. (2000). Robust Logistic Regression for Binomial Responses. Available at SSRN: https://ssrn.com/abstract=1763301 or http://dx.doi.org/10.2139/ssrn.1763301

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals Statistics,* **15,** 692–656. MR0888431

Warwick, J. and Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, **75**, 581–588. MR2162547

Zelterman, D. (2005). *Models for Discrete Data*. Oxford University Press, New York. MR1707334