KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

# Robust estimation for ordinal regression

C. Croux, G. Haesbroeck and C. Ruwet

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1110

# Robust estimation for ordinal regression

C. Croux[a,*], G. Haesbroeck[b], C. Ruwet[b]

[a] *KULeuven, Faculty of Business and Economics, Leuven, Belgium*
[b] *University of Liege, Department of Mathematics, Liege, Belgium*

**Abstract**

Ordinal regression is used for modelling an ordinal response variable as a function of some explanatory variables. The classical technique for estimating the unknown parameters of this model is Maximum Likelihood (ML). The lack of robustness of this estimator is formally shown by deriving its breakdown point and its influence function. To robustify the procedure, a weighting step is added to the Maximum Likelihood estimator, yielding an estimator with bounded influence function. We also show that the loss in efficiency due to the weighting step remains limited. A diagnostic plot based on the Weighted Maximum Likelihood estimator allows to detect outliers of different types in a single plot.

*Keywords:* Breakdown point, Diagnostic plot, Influence function, Ordinal regression, Weighted Maximum Likelihood, Robust distances.

## 1. Introduction

Logistic regression is frequently used for classifying observations into two groups. When dealing with more than two groups, this model needs to be

*Corresponding author
*Email address:* Christophe.Croux@econ.kuleuven.be (C. Croux)

generalized. When the labels of these groups are naturally ordered, an ordinal regression model can be fitted to the data. The group label is then the ordinal response variable. Ordinal variables occur frequently in practice, e.g. in surveys where respondents have to specify whether they strongly disagree, disagree, are indifferent, agree or strongly agree with a given statement. As illustration in this paper, we use the wine data set of Bastien et al. (2005). These data characterize the quality of 34 years of Bordeaux wine, the quality being assessed on the ordinal scale Poor-Average-Good. The quality is assumed to be related to four explanatory variables: temperature (measured by the sum of average day temperatures in degrees celsius), sunshine (duration of sunshine in hours), heat (number of very warm days) and rain (rain height in mm). In Figure 1, two scatter plots representing the data are given. One can see that the explanatory variables contain relevant information to characterize the quality of the wine. For example, rainy years correspond generally to poor wines as well as years with few sunshine and warm days.

Following Anderson and Philips (1981), we introduce the ordinal regression model of interest via a latent, unobservable continuous variable $Y^*$. This latent variable depends on a vector of explanatory variables $X = (X_1, \ldots, X_p)^t$ as

$$Y^* = \beta^t X + \varepsilon, \tag{1}$$

where $\beta$ is a $p$-vector of unknown regression parameters and $\varepsilon$ is a random variable with cumulative distribution function $F$. The observed ordinal variable $Y$ takes as values the labels $1, \ldots, J$. We have

$$Y = j \text{ if } \alpha_{j-1} < Y^* \leq \alpha_j, \text{ for } j = 1, \ldots, J, \tag{2}$$

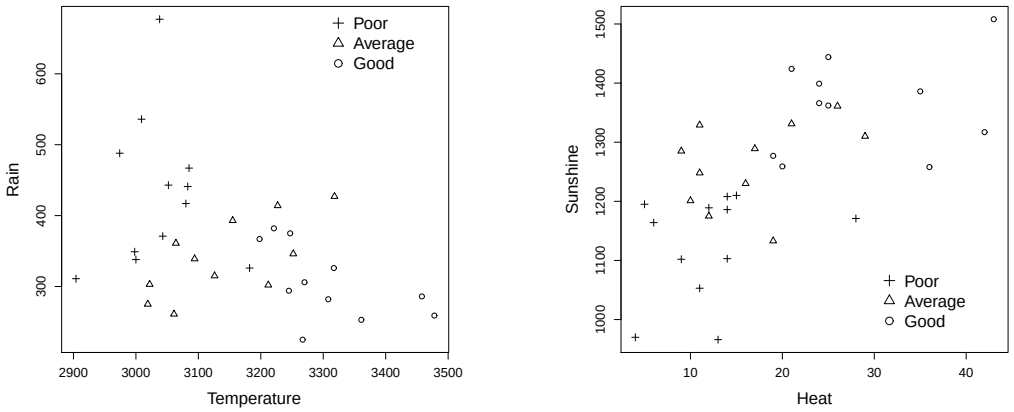


Figure 1: Scatter plots of Rain versus Temperature (left panel) and Sunshine versus Heat (right panel) for the Wine data categorized as Poor, Average or Good. The value of the dependent variable "quality of Wine" is represented by the corresponding symbol.

where the $\alpha_j$ are unobserved thresholds with $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_{J-1} < \alpha_J = \infty$. Combining (1) and (2) yields

$$\mathbb{P}[Y = j|X = x] = F\left(\alpha_j - \beta^t x\right) - F\left(\alpha_{j-1} - \beta^t x\right) \quad \text{for } j = 1, \ldots, J. \tag{3}$$

We assume that $F$ is strictly increasing and symmetric around zero, so $F(0) = 0.5$. Standard choices for the distribution function $F$ are the logistic link function, $F(t) = 1/(1 + e^{-t})$, corresponding to the logistic distribution, or the probit link function, $F(t) = \Phi(t)$, with $\Phi$ the cdf of the standard normal distribution.

Fitting an ordinal regression model requires the estimation of $J-1+p$ parameters, i.e. the $J-1$ thresholds $\alpha = (\alpha_1, \ldots, \alpha_{J-1})^t$ and the $p$ components of $\beta$. Maximizing the log-likelihood function is the most common procedure to obtain these estimations (e.g. Anderson and Philips, 1981; Franses and Paap, 2001; Powers and Xie, 2008).

In general, when estimating parameters by means of Maximum Likelihood, it is expected that outliers will have a devastating impact on the results. The lack of robustness of the Maximum Likelihood method in the logistic regression model has been already extensively studied in the literature, e.g. breakdown points have been computed (e.g. Croux et al., 2002; Müller and Neykov, 2003) and influence functions have been derived (Croux and Haesbroeck, 2003). Also, robust estimation techniques have been introduced (e.g. Carroll and Pederson, 1993; Wang and Carroll, 1995; Bianco and Yohai, 1996; Gervini, 2005; Bondell, 2008; Hobza et al., 2008; Hosseinian and Morgenthaler, 2011). However, to our best knowledge, similar results are not yet available for the ordinal regression model.

In this paper, we investigate the lack of robustness of the ML procedure in the ordinal regression setting by computing breakdown points and influence functions. A robust alternative consisting of a weighting step added to the ML methodology is then presented. Section 2 defines the Maximum Likelihood estimator and states the conditions under which this estimator exists. In Section 3, the breakdown point of the ML estimator is derived and shown to go to zero as the sample size tends to infinity. A robust alternative is introduced in Section 4. In Section 5, the influence functions of the classical and robust estimators are computed and they are then used in Section 6 to construct a diagnostic plot detecting influential points. The statistical precision of the robust estimators is discussed in Section 7. Finally, Section 8 makes some conclusions.

## 2. Maximum likelihood estimation

### *2.1. Definition*

Let $Z_n = \{(x_i, y_i) : i = 1, \ldots, n\}$ be a sample of size $n$ where the vector $x_i \in \mathbb{R}^p$ contains the observed values of the $p$ explanatory variables and $y_i$, with $1 \leq y_i \leq J$, indicates the membership to one of the $J$ groups. The Maximum Likelihood estimator is obtained by maximizing the log-likelihood function, i.e.

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)\in\mathbb{R}^{J-1+p}}{\operatorname{argmax}} \; l(\alpha, \beta), \tag{4}$$

under the constraint $\alpha_1 < \ldots < \alpha_{J-1}$, with

$$l(\alpha, \beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_{ij} \log\left(F\left(\alpha_j - \beta^t x_i\right) - F\left(\alpha_{j-1} - \beta^t x_i\right)\right), \tag{5}$$

and where the indicator function $\delta_{ij}$ takes the value 1 when $y_i = j$ and 0 otherwise.

To explicitly take into account the ordering constraint in the maximization, Franses and Paap (2001) recommend to re-parameterize the log-likelihood function by replacing the vector of thresholds $\alpha$ by $\gamma = (\gamma_1, \ldots, \gamma_{J-1})^t$ defined as

$$\begin{aligned} \alpha_1 &= \gamma_1 \\ \alpha_j &= \gamma_1 + \sum_{k=2}^{j} \gamma_k^2, \text{ for } j = 2, \ldots, J-1. \end{aligned} \tag{6}$$

The parameter $\gamma$ is uniquely identified by asking that $\gamma_j \geq 0$, for $j > 1$. The

log-likelihood function (5) can be rewritten as

$$l(\gamma,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{J}\delta_{ij}\log\left(F(\gamma_1+\sum_{k=2}^{j}\gamma_k^2-\beta^t x_i)-F(\gamma_1+\sum_{k=2}^{j-1}\gamma_k^2-\beta^t x_i)\right)$$
$$= \sum_{i=1}^{n}\varphi(x_i,y_i,(\gamma,\beta)). \tag{7}$$

The Maximum Likelihood estimators of $\gamma$ and $\beta$ are then given by

$$(\hat{\gamma},\hat{\beta}) = \underset{(\gamma,\beta)\in\mathbb{R}^{J-1+p}}{\operatorname{argmax}} \; l(\gamma,\beta). \tag{8}$$

The advantage of the optimization problem in (8) is that no constraints need to be put on the parameters: the resulting estimates for the thresholds $\alpha$ will be automatically ordered. Furthermore, equality of two thresholds implies a zero values for some $\gamma_j$, with $j > 1$, yielding minus infinity for the objective functions in (7), and they can be excluded from the solutions set.

### *2.2. Existence*

When working with a binary regression model, it is well known that an overlap between the two groups of observations is necessary and sufficient for existence and uniqueness of the Maximum Likelihood estimates (Albert and Anderson, 1984). For ordinal regression, the existence of the ML estimates has also been characterized by overlap conditions. Haberman (1980) states these conditions in algebraic terms. Proposition 1 below summarizes these conditions.

**Proposition 1** (Haberman, 1980)**.** *Let the set $I^j$ contain the indexes of the observations for which $y_i = j$, for $j = 1,\ldots,J$. The Maximum Likelihood estimate exists if and only if*

*(i)* $\sum_i \delta_{ij} \geq 1$ *for all* $j = 1, \ldots, J$

*(ii) For all* $\alpha$ *and* $\beta$*, there exists an index* $j$ *in* $\{1, \ldots, J\}$*, and there exists an* $i$ *in* $I^j$ *such that* $x_i^t\beta < \alpha_{j-1}$ *or* $x_i^t\beta > \alpha_j$*.*

In words, these conditions state that the ML estimates exist if and only if $(i)$ there is no empty group and $(ii)$ there is overlap between at least two groups with consecutive labels. As only one overlap is necessary to ensure existence and uniqueness, the overlap condition is not more stringent than in the binary case.

## 3. Breakdown point of the Maximum Likelihood estimator

In the logistic regression setting, Croux et al. (2002) showed that the ML estimator never explodes when outliers are added to the data. On the other hand, the estimated slope goes to zero when adding $2p$ well chosen outliers. This behavior also holds in ordinal regression, as Propositions 2 and 3 below prove.

Let $z_i = (x_i, y_i)$ denote the $i$th observation and $Z'_{n+m} = \{z_1, \ldots, z_n, z_{n+1}, \ldots, z_{n+m}\}$ the initial sample with $m$ outliers, $z_{n+1}, \ldots, z_{n+m}$, added. The ML estimator of the thresholds and slope parameter is denoted by $\hat{\theta}(Z_n) = \hat{\theta}_n$. The notation $\|.\|$ refers to the Euclidean norm.

**Definition.** The *explosion breakdown point* of $\hat{\theta}_n$ at the sample $Z_n$ is the minimal fraction of outliers that needs to be added to the initial sample to get the estimator over all bounds, i.e.

$$\varepsilon^+(\hat{\theta}_n, Z_n) = \frac{m^+}{n + m^+} \text{ with } m^+ = \min \left\{ m \in \mathbb{N}_0 \,\middle|\, \sup_{z_{n+1}, \ldots, z_{n+m}} \|\hat{\theta}(Z'_{n+m})\| = \infty \right\}.$$

The above definition is the addition breakdown point of an estimator. Alternatively, once could consider the replacement breakdown point, where observations are replaced by outliers until the estimator goes over all bounds. The replacement breakdown point is less appealing in our setting, since one could make the overlap condition of proposition 1 fail, causing a breakdown of the estimator due to its non-existence. When adding outliers to the data, the overlap condition remains verified.

Assuming that overlap holds initially, Proposition 2 formally proves that the ML estimator in ordinal regression is uniformly bounded above when adding an arbitrary number of outliers in the data. The proof is given in the Appendix.

**Proposition 2.** *Assume that* $\|\hat{\theta}(Z_n)\| < \infty$*, with* $\hat{\theta}(Z_n)$ *the Maximum Likelihood estimator computed on the sample* $Z_n$*. Then* $\varepsilon^+(\hat{\theta}_n, Z_n) = 1$*.*

While Proposition 2 shows that the explosion breakdown point of the ML estimator is 100%, Proposition 3 derives an upper bound for the breakdown point of the slope estimator taking both implosion and explosion behaviors into account.

**Definition.** The *breakdown point* of $\hat{\beta}_n$ at the sample $Z_n$ is the minimal fraction of outliers that needs to be added to the initial sample in order to make the estimator tend to zero or to infinity:

$$\varepsilon(\hat{\beta}_n, Z_n) = \frac{m^*}{n + m^*}$$

with $m^* = \min(m^+, m^-)$ and

$$m^+ = \min \left\{ m \in \mathbb{N}_0 \left| \sup_{z_{n+1}, \ldots, z_{n+m}} \|\hat{\beta}(Z'_{n+m})\| = \infty \right. \right\}$$

$$m^- = \min \left\{ m \in \mathbb{N}_0 \left| \inf_{z_{n+1},\ldots,z_{n+m}} \|\hat{\beta}(Z'_{n+m})\| = 0 \right. \right\}.$$

**Proposition 3.** *At any sample $Z_n$, the breakdown point of the slope of the ML estimator for the ordinal regression model satisfies*

$$\varepsilon(\hat{\beta}_n, Z_n) \leq \frac{pJ}{n + pJ},$$

*where $p$ is the number of explanatory variables and $J$ the number of groups.*

Proposition 3 shows that the ML estimator is not robust since its asymptotic breakdown point, i.e. $\lim_{n\to+\infty} \varepsilon(\hat{\beta}_n, Z_n)$, is zero. This lack of robustness does not come from an explosion but rather from an implosion of the slope estimator toward zero. It is interesting to note that the implosion of the slope estimate does not imply implosion of the estimations of the thresholds. Indeed, as $\hat{\beta}$ goes to 0, $\hat{\alpha}_j$ tends to $F^{-1}(p_1 + \ldots + p_j)$, where $p_j$ is the frequency of observations in the $j$th group. In general, this limit does not vanish to zero.

To illustrate the implosion breakdown of the slope estimator, let us consider the simulated data set pictured in Figure 2, with $n = 50$ observations classified into $J = 3$ ordered groups. The group membership depends on $p = 2$ explanatory variables (simulated as independent standard normals) while the error term is distributed according to the logistic distribution. The true values of the parameters are set at $\beta = (-1, 1.5)^t$ and $\alpha_1 = -\alpha_2 = -1$. This choice of thresholds leads to three groups of equivalent size. Based on this initial data set, the ML estimate of $\beta$ is given by $\hat{\beta} = (-1.30, 1.60)^t$, yielding $\|\hat{\beta}\| = 2.06$. The ML estimator yields a misclassification rate (which is the proportion of observations for which $\hat{y} \neq y$) equal to 28%.

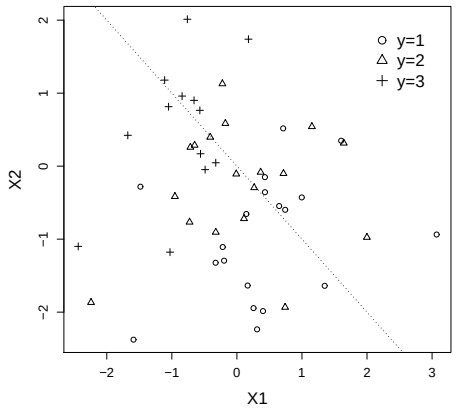


Figure 2: Scatter plot of the simulated data set. The dotted line represents the shift of the additional observation $((s, -s), 3)$.

Both the estimator and the misclassification rate may be completely perturbed by the introduction of a single outlier in the data. Add one observation with $x = (s, -s)$ moving along the dotted line in Figure 2 and with $y = 3$. This additional observation is most outlying when $s$ is positive since it lies then in the region of observations associated with the smallest possible $y$-score. When $s$ is negative, the observation is outlying in the space of explanatory variable (if $s$ is large) but it has the expected value as far as the $y$-variable is concerned.

For each value of $s$, the parameters of the ordinal regression model are estimated by Maximum Likelihood and the corresponding misclassification rate is computed. The resulting norm of $\hat{\beta}$ and the rate of misclassification are represented in Figure 3 with respect to $s$. As expected, negative values of $s$ yield an additional observation which does not bias too much the estimator. On the other hand, as soon as $s$ gets positive, the impact of the added

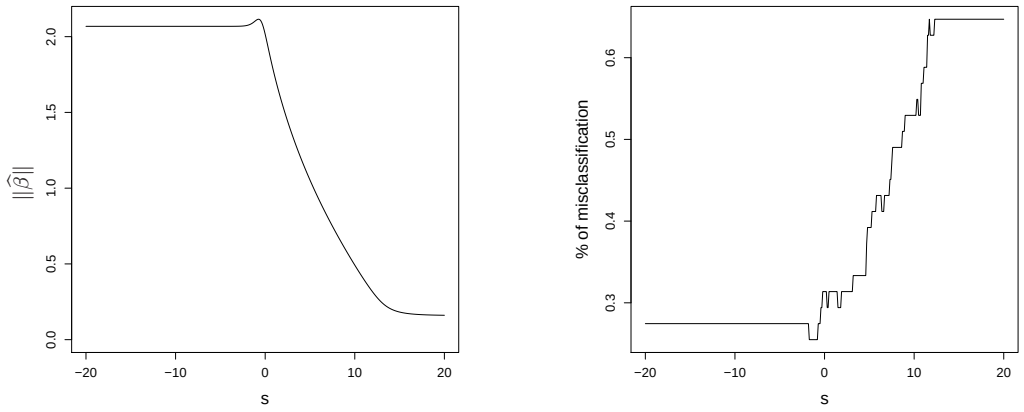


Figure 3: Simulated data set with $((s, -s), 3)$ as additional observation. Left panel: Norm of the slope parameter. Right panel: Proportion of misclassified observations.

observation becomes apparent and gets even quite extreme as $s$ increases. Not only the norm of the slope estimator goes to zero but the misclassification rate reaches a limit of about 66%, close to the classification performance of a random guess.

## 4. Weighted Maximum Likelihood Estimator

In this Section, we construct a robust alternative to the ML estimator. The most simple way to decrease the influence of outliers is to downweight them by adding weights into the log-likelihood function. As such, a weighted Maximum Likelihood estimator is obtained, as already suggested and studied in the simple logistic regression setting (e.g. Carroll and Pederson, 1993; Croux and Haesbroeck, 2003; Croux et al., 2008).

Many different types of weights could be designed. Here, the downweighting will be done using robust distances computed in the space of the ex-

planatory variables. Let $m$ and $S$ denote robust estimates of the location and covariance matrix based on $x_1, \ldots, x_n$. In this paper, the Minimum Covariance Determinant estimator (Rousseeuw and Van Driessen, 1999) with a breakdown point of 25% has been used. The weight $w_i$ attributed to the $i$th observation is then given by

$$w_i = W(d_i) \text{ with } d_i = (x_i - m)^t S^{-1}(x_i - m), \tag{9}$$

for a given weight function $W$. The robust distances measure the outlyingness of each observation $x_i$, for $i = 1, \ldots, n$. They are only computed from the continuous explanatory variables. In particular, dummy variables are not taken into account when computing the distances.

The Weighted Maximum Likelihood (WML) estimator is then the solution of the following maximization problem

$$(\hat{\gamma}, \hat{\beta}) = \underset{(\gamma,\beta)\in\mathbb{R}^{J-1+p}}{\operatorname{argmax}} \sum_{i=1}^{n} w_i \varphi(x_i, y_i, (\gamma, \beta)), \tag{10}$$

where $\varphi$ is the log-likelihood of an individual observations, as in equation (7).

Obviously, if one takes a constant weight function, the usual ML estimator is obtained again. A typical weight function is the step function $W_{0/1}(d) = I(d \leq \chi^2_p(0.975))$ where $\chi^2(0.975)$ is the 97.5% quantile of the chi-squared distribution with $p$ degrees of freedom. With such a weight function, observations lying far away from the bulk of the data are discarded. As only extreme observations are discarded, one expects that in most cases the overlap condition of Proposition 1 will still hold. Nevertheless, to guarantee existence of overlap, a smoother weight function is prefered, as the Student weight function defined by $W_\nu(d) = (p + \nu)/(d + \nu)$, where $\nu$ is the degree

of freedom. The larger $\nu$, the less observations are downweigthed. We take $\nu = 3$. Far away observations do get small weights but are not discarded completely from the data set. Therefore, the weighted ML estimate using the $W_3$ weight function exists as soon as the ML estimate exists.

To illustrate numerically the robustness of the WML estimator, we repeat the experiment discussed in Section 3. Similar as Figure 3, Figure 4 shows how the norm of a robust estimate of the slope parameter depends on the position of a single added outlier with value $x = (s, -s)^t$ and $y = 3$. It can be seen that the norm of the slope does not tend to zero anymore, and is only changing in the region where the added observation is not too different from the bulk of the data. We see that the curve corresponding to the Student weight function is smoother than the one based on the 0/1 weights, as expected. Figure 4 also shows that the misclassification rate is only slightly varying with the values of the outlier, and nowhere close to 66%, the misclassification rate corresponding to random guessing.

## 5. Influence function

### *5.1. Derivation*

Breakdown points are global measures of robustness measuring robustness in presence of large amounts of contamination. On the other hand, influence functions are local measures of robustness, characterizing the impact of an infinitesimal proportion of contamination located at a point $(x, y)$. The influence function of a statistical functional $T$ at the model distribution $H_0$ is defined by

$$\mathrm{IF}((x, y); T, H_0) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)H_0 + \varepsilon \Delta_{(x,y)}) - T(H_0)}{\varepsilon} \qquad (11)$$

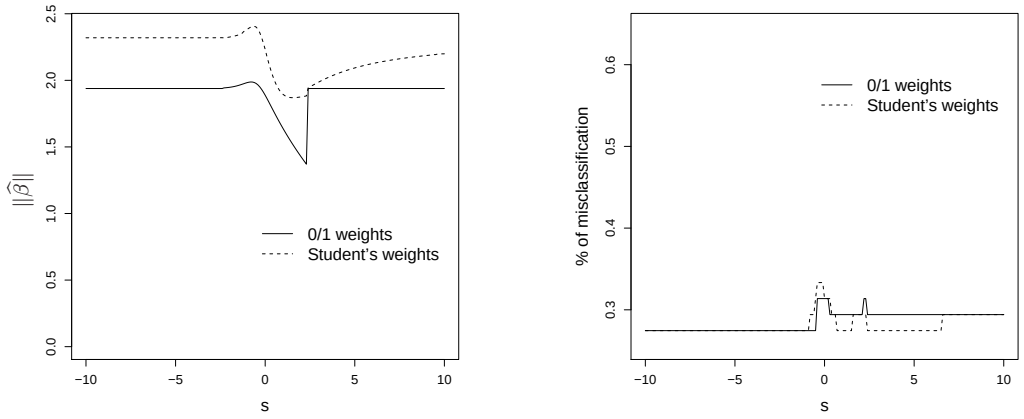


Figure 4: Simulated data set with $((s, -s), 3)$ as additional observation. Left panel: Norm of the slope parameter. Right panel: Proportion of misclassified observations. Parameters are estimated with WML using 0/1 weights (solid line) or Student's weights (dashed line).

where $\Delta_{(x,y)}$ is the Dirac distribution having all its mass at $(x, y)$ for given $x \in I\!R^p$ and $y \in \{1, \ldots, J\}$. Contamination on $y$ is restricted to its possible values since any other choice for $y$ would be easily detected.

We first derive the influence function of the statistical functionals related to the estimation of the parameters $\gamma$ and $\beta$. They are the solutions of an unconstrained problem, see equation (8), and easier to deal with. Let $\theta$ represent the joint vector $(\gamma, \beta)$. The statistical functional relative to the ML estimator in ordinal regression is given by

$$\theta_{ML}(H) = \underset{\theta \in \mathbb{R}^{J-1+p}}{\operatorname{argmax}} E_H[\varphi(X, Y; \theta)],$$

for any distribution $H$ of the variables $(X, Y)$. The first order condition yields

$$E_H[\psi(X, Y; \theta)] = 0 \text{ with } \psi(X, Y; \theta) = \frac{\partial \varphi(X, Y; \theta)}{\partial \theta}.$$

At the model distribution $H_0$, equation (3) holds, and it follows that $\theta_{ML}(H_0) = \theta_0$, with $\theta_0$ the true parameter vector. We observe that $\theta_{ML}$ is simply a M-type statistical functional for which the IF is readily available (Hampel et al., 1986, page 230) and given by:

$$\mathrm{IF}((x,y);\theta_{ML},H_0) = -E_{H_0}\left[\left.\frac{\partial^2\varphi(X,Y;\theta)}{\partial\theta^2}\right|_{\theta_0}\right]^{-1}\psi(x,y;\theta_0). \qquad (12)$$

The first factor in (12) is a constant $(J-1+p)$-square matrix, which we demote by $M(\psi,H_0)$, independent of $x$ and $y$. Its explicit form is given in the Appendix. The shape of the influence function is mainly determined by the second factor $\psi$ with components $(\psi_1,\ldots,\psi_{J-1},\psi_\beta)$, and defined as

$$\begin{aligned}
\psi_1(x,y;\theta) &= \sum_{j=1}^{J} I(y=j)\frac{f(\alpha_j-\beta^t x)-f(\alpha_{j-1}-\beta^t x)}{F(\alpha_j-\beta^t x)-F(\alpha_{j-1}-\beta^t x)},\\
\psi_k(x,y;\theta) &= 2\gamma_k\left[I(y=k)\frac{f(\alpha_k-\beta^t x)}{F(\alpha_k-\beta^t x)-F(\alpha_{k-1}-\beta^t x)}\right.\\
&\qquad\left.+\sum_{j=k+1}^{J} I(y=j)\frac{f(\alpha_j-\beta^t x)-f(\alpha_{j-1}-\beta^t x)}{F(\alpha_j-\beta^t x)-F(\alpha_{j-1}-\beta^t x)}\right]
\end{aligned}$$

for $k=2,\ldots,J-1$ and

$$\psi_\beta(x,y;\theta) = -x\sum_{j=1}^{J} I(y=j)\frac{f(\alpha_j-\beta^t x)-f(\alpha_{j-1}-\beta^t x)}{F(\alpha_j-\beta^t x)-F(\alpha_{j-1}-\beta^t x)},$$

with $f=F'$ the density function of the error term in (1).

It is easy to see that the influence function of ML is bounded in $y$, since $y$ only enters the IF through the indicator functions $I(y=j)$. The first $J-1$ components of $\psi$ are also bounded in $x$, as $f$ and $F$ are bounded functions (at least for the logit and probit link functions). The slope component of $\psi$, however, is unbounded in the value of the covariate $x$, proving the lack of

robustness of the ML-estimator in presence of small amounts of contamination.

Let us now turn to the derivation of the influence function of the WML statistical functional $\theta_{WML}$. It is defined as

$$\theta_{WML}(H) = \underset{\theta \in \mathbb{R}^{J-1+p}}{\operatorname{argmax}} E_H[W(D_H(X))\psi(X, Y; \theta)]$$

where $(X, Y) \sim H$. With $G_H$ the marginal distribution of $X$, the distance function $D_H$ is given by $D_H(x) = (x - \mu(G_H))^t \Sigma(G_H)^{-1}(x - \mu(G_H))$ where $(\mu(G_H), \Sigma(G_H))$ are the location and covariance functionals corresponding to the MCD estimator.

Using the explicit expression of $\psi$ given above, it is easy to check that conditional Fisher consistency holds, i.e. $E_{H_0}[\psi(X, Y; \theta_0)|X = x] = 0$ for all $x \in \mathbb{R}^p$. Lemma 1 of Croux and Haesbroeck (2003) leads then to the following expression for the influence function, $\mathrm{IF}((x, y); \theta_{WML}, H_0)$, of the Weighted Maximum Likelihood estimator:

$$-E_{H_0}\left[W(D_{H_0}(X))\left.\frac{\partial^2 \varphi(X, Y; \theta)}{\partial \theta^2}\right|_{\theta_0}\right]^{-1} W(D_{H_0}(x))\psi(x, y; \theta_0). \qquad (13)$$

Leaving aside the constant matrix, the impact of an infinitesimal contamination at $(x, y)$ on WML is measured by means of the same function $\psi$ as before, now multiplied by the weight function. The influence function of WML remains therefore bounded in $y$ but also becomes bounded with respect to outlying values of the explanatory variables as soon as $x\, W(D_{H_0}(x))$ is bounded. Both for the Student weight function and for the 0/1 weight function we get bounded influence functions.

From expressions (12) and (13), the influence functions of the functionals corresponding to the thresholds $\alpha_1, \ldots, \alpha_{J-1}$, are readily obtained. Denote $A_1, \ldots, A_{J-1}$ the statistical functionals corresponding to the estimators of the components of the parameter $\alpha$, and $C_1, \ldots, C_{J-1}$ the first $J-1$ components of the statistical functional $\theta_{ML}$ or $\theta_{WML}$. Using definition (6), one gets

$$\mathrm{IF}\left((x,y)\,;A_1,H_0\right) = \mathrm{IF}\left((x,y)\,;C_1,H_0\right)$$

$$\mathrm{IF}\left((x,y)\,;A_j,H_0\right) = \mathrm{IF}\left((x,y)\,;C_1,H_0\right) + 2\sum_{k=2}^{j}\gamma_k \mathrm{IF}\left((x,y)\,;C_k,H_0\right)$$

for $j = 2, \ldots, J-1$.

*5.2. Numerical Illustrations*

Let us look at some graphical representations of the influence functions. We take for $F$ the probit link but similar results hold for the logit link. We consider the model $Y^* = \beta^t X + \varepsilon$, where the covariates $X$ follow a standard normal distribution. The ordinal variable $Y$ is then given by (2), where the thresholds are such that every group has equal probability of occurrence.

We focus here on the univariate case, $p = 1$, with $J = 3$ groups and we set $\beta = 1.5$. Figure 5 shows the influence functions of $A_1$ (upper panels) and of the functional $B$ estimating the slope parameter $\beta$ (lower panels) for the ML estimator (left panels) and for the WML estimator based on Student's weights with $\nu = 3$ (right panels), as a function of $x$. Several curves are drawn, one for each possible value of $y$. While the influence functions of the ML estimator are unbounded, adding weights yields bounded influence functions for the WML estimator. Bounded IF are also obtained for the WML estimator with the 0/1 weighting scheme (not reported) but jumps appear due to the discontinuity of the step weight function.

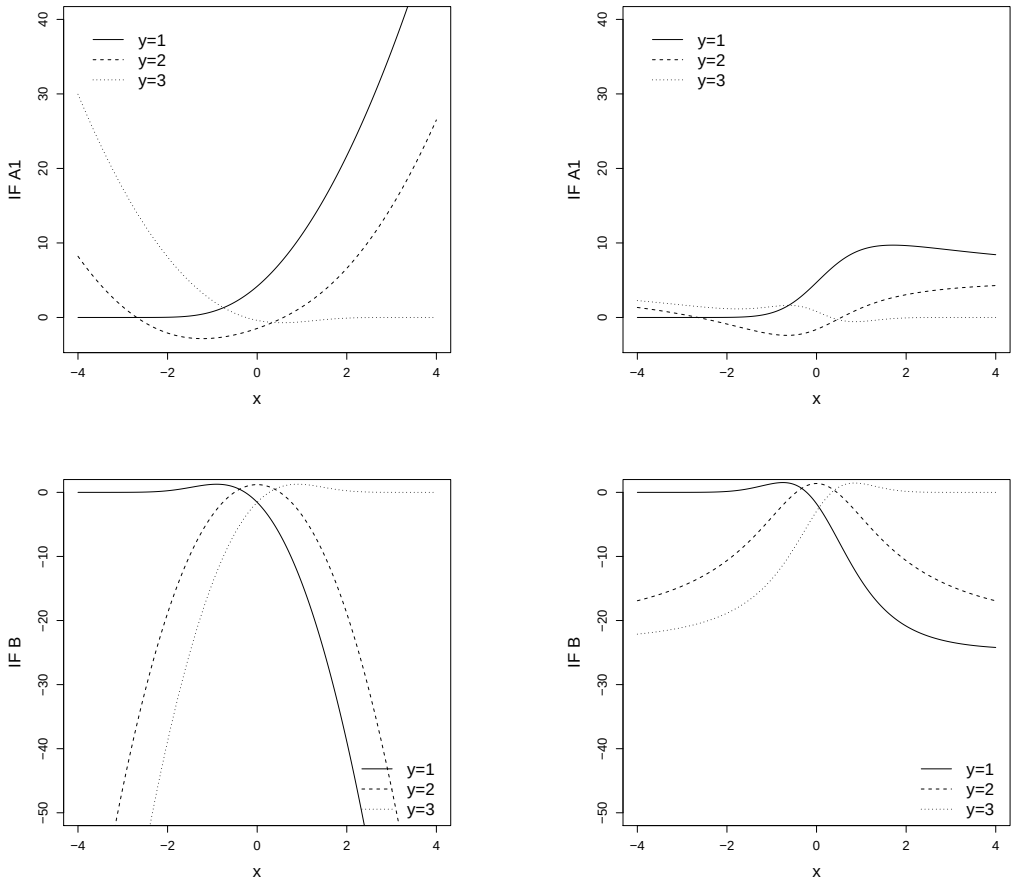


Figure 5: Influence functions for the first threshold $A_1$ (upper row) and for the slope $B$ (lower row) as functions of $x$ with $y \in \{1, 2, 3\}$, for $p = 1$ and with $J = 3$ groups. Left panels: ML estimator. Right panels: WML estimator based on Student weight function with $\nu = 3$.

It is worth to interpret in more detail these influence functions. First, as $\beta$ is positive, large negative values of the explanatory variable yield a fitted value of the ordinal variable, $\hat{y}$, equal to its smallest score ($\hat{y} = 1$) while large positive ones would correspond to the highest score ($\hat{y} = 3$). Looking now at the influence function of the first threshold (upper plots of Figure 5), one can observe that negative (resp. positive) $x$ values have a zero influence when associated with $y = 1$ (resp. $y = 3$), showing that these points are not influential even on the ML estimator. The influence is smaller (in absolute value) for those $x$ lying in the area corresponding to $\hat{y} = y$ than elsewhere. The same remarks hold for the IF of the slope parameter (see lower panel of Figure 5). These influence functions also shows the expected symmetry.

This leads to the definition of several types of outliers in the ordinal regression setting. For outlying values of $x$, i.e. for leverage points, some couples $(x, y)$ are influential, while others are not. If $\hat{y} = y$, the corresponding outlier has less influence on the estimation of the parameters. It can be labelled as a *good leverage point*, as in linear regression (Rousseeuw and Leroy, 1987). On the other hand, when $\hat{y} \neq y$, this outlier might be highly influential and can be considered as a *bad leverage point*. It may happen that $\hat{y} \neq y$ even if $x$ is not outlying in the space of the explanatory variables. In that case, we talk about *vertical outliers*.

## 6. Diagnostic plot

It has been shown that atypical observations may have an important effect on the Maximum Likelihood estimator. In order to detect the potentially influential observations beforehand, diagnostic measures could be computed.

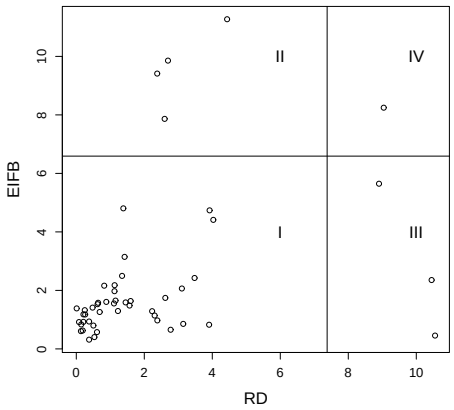


Figure 6: Illustration of the diagnostic plot detecting influential points (vertical axis) and leverage points (horizontal axis). The influence measure is plotted versus the robust distance (RD).

Here follow the approach of Pison and Van Aelst (2004), based on influence functions.

The influence of each observation on the classical estimator is measured and plotted with respect to the robust distances (RD) computed on the continuous explanatory variables in the data set. Figure 6 displays such a diagnostic plot. The vertical and horizontal lines correspond to cutoff values: a distance or influence measure larger than the cutoff value indicates an atypical observation. As shown on Figure 6, we get four parts: Part I contains the regular observations, Part II corresponds to the vertical outliers, part III to the good leverage points and Part IV to the bad leverage points.

To compute the influence measures, we evaluate the influence function of the ML estimator at each observed couple $(x_i, y_i)$. As expression (12) shows, this IF still depends on unknown quantities: the model distribution of the

explanatory variables, $G$, and the true values of the parameters, $\theta_0$. To avoid the masking effect, Pison and Van Aelst (2004) suggest to estimate $G$ and $\theta_0$ in a robust way. The parameter $\theta_0$ is estimated by WML. The distribution $G$ is estimated by the empirical distribution based on the observations which are not detected as outliers in the space of explanatory variables, i.e. for which $d_i \leq \chi^2_p(0.975)$. Recall that $d_i$ are the robust distances defined in (9). Since $G$ and $\theta_0$ are replaced by estimates, we speak about *empirical influence functions* (EIF). The overall influence measures for the threshold and regression slope estimators are then

$$\text{EIFA}_i = \frac{\|\text{EIF}((x_i, y_i), A)\|}{\sqrt{J-1}} \text{ and } \text{EIFB}_i = \frac{\|\text{EIF}((x_i, y_i), B)\|}{\sqrt{p}}$$

where the factors $1/\sqrt{J-1}$ and $1/\sqrt{p}$ scale the norms. As cutoff for the influence measure, the empirical 95% quantile of a set of simulated overall empirical influence functions is chosen, as in Pison and Van Aelst (2004).

Two examples will illustrate the usefulness of this diagnostic plot in practice. The first one is based again on the simulated data set of Section 3. In this simple and simulated setting, the different types of outliers may be easily detected by visual inspection of the data. We will show that the same detection may be obtained via the corresponding diagnostic plot.

The left panel of Figure 7 displays the data together with the robust fit obtained with the Weighted Maximum Likelihood estimator. The dashed lines separate the plane in three parts, according to the fitted value of the respons variable. The diagnostic plot for the slope parameter is shown on the right panel (similar results hold for the thresholds). Some particular observations (numbered from 1 to 8) are pinpointed on these two plots. Observations 1, 2, 3 and 4 are leverage points. Out of these four observations, only one

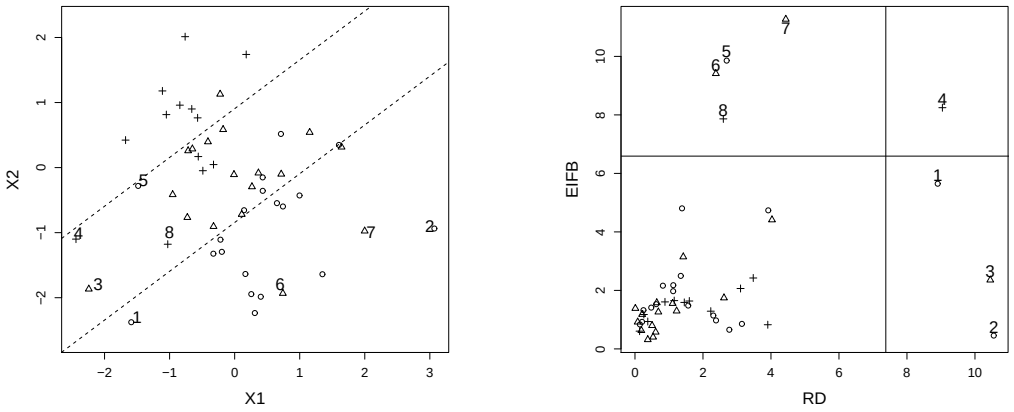


Figure 7: Original simulated data set together with the fitted separating lines based on the WML estimator with 0/1 weights (left panel) and corresponding diagnostic plot (right panel).

(observation 4) is a bad leverage point. Observations labelled as 5, 6, 7 and 8 are not outlying in the space of explanatory variables but are misclassified. They are vertical outliers and lie indeed in Part II of the diagnostic plot.

For the second illustration, let us come back to the wine data set presented in the Introduction. There are $n = 34$ observations and $p = 4$ explanatory variables; visual analysis of the data is no longer possible. Figure 8 gives the diagnostic plot based on the slope parameter. Several observations (numbered by their index which gives the corresponding year of the production of the wine) lie in the outlying parts of the plot. Eight observations are outlying in the space of explanatory variables. Two of them are bad leverage points (years 1928 and 1956) while the others (years 1927, 1929, 1932, 1935, 1947 and 1949) are not influential. Except for 1935, these years are known to be either disastrous or exceptional as far as their climatic conditions are

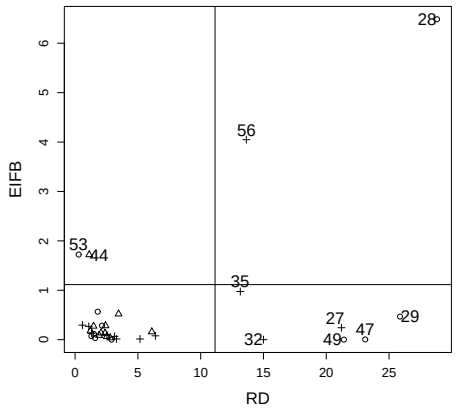


Figure 8: Diagnostic plot for the Wine data set.

concerned. Years 1944 and 1953 are flagged as vertical outliers. They correspond indeed to wines for which the observed quality is not validated by the estimated model.

## 7. Simulation study

While adding weights in the log-likelihood function makes the ML estimator robust, it also leads to a loss in statistical efficiency. By means of a modest simulation study, we show that this loss remains limited. For $m = 5000$ samples of size $n = 50$ or $n = 200$, observations were generated according to the ordinal regression model, with $F$ the probit link function, and the covariates following a $N(0, I_p)$ distribution, for $p = 2, 3$ and 5. The thresholds $\alpha = (\alpha_1, \ldots, \alpha_{J-1})^t$ are selected such that every value of the ordinal variable $Y$ has the same probability to occur. We present results for $J = 3$ groups. The slope parameter is taken as $\beta = (1, 1, \ldots, 1)^t$. For every generated sample, the parameters are estimated with ML and WML using

| | | $p=2$ | | | $p=3$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | 50 | 200 | $\infty$ | 50 | 200 | $\infty$ | 50 | 200 | $\infty$ |
| $W_{0/1}$ | $\text{Eff}_n(\alpha)$ | .984 | .989 | .991 | .951 | .986 | .990 | .767 | .973 | .986 |
| | $\text{Eff}_n(\beta)$ | .941 | .961 | .963 | .922 | .962 | .965 | .686 | .951 | .964 |
| $W_3$ | $\text{Eff}_n(\alpha)$ | .953 | .947 | .943 | .940 | .926 | .921 | .975 | .904 | .893 |
| | $\text{Eff}_n(\beta)$ | .959 | .933 | .922 | .956 | .929 | .909 | .978 | .919 | .894 |

Table 1: Relative efficiencies of WML w.r.t. ML for the threshold and slope estimators, using 0/1 weights ($W_{0/1}$) and the Student weight function ($W_3$).

both the 0/1 and the student weight functions.

For every component of $\alpha$ and $\beta$, we compute the Mean Squared Errors (MSE) of the estimators, and summarize the relative finite-sample efficiencies of the Weighted ML versus the ML estimator as

$$\text{Eff}_n(\alpha) = \frac{1}{J-1}\sum_{k=1}^{J-1}\frac{\text{MSE}(\hat{\alpha}_{k,ML})}{\text{MSE}(\hat{\alpha}_{k,WML})} \text{ and } \text{Eff}_n(\beta) = \frac{1}{p}\sum_{k=1}^{p}\frac{\text{MSE}(\hat{\beta}_{k,ML})}{\text{MSE}(\hat{\beta}_{k,WML})}.$$

The results are reported in Table 1. In the column $n = \infty$, we report asymptotic efficiencies computed using the rule

$$\text{ASV}(T, H_0) = E_{H_0}\left[\text{IF}^2((X,Y); T, H_0)\right],$$

with $T$ the functional corresponding to the estimation of one component of $\alpha$ or $\beta$ (Hampel et al., 1986, page 85). Using numerical integration and the expression for the influence functions derived in Section 5, we obtain the asymptotic variances and efficiencies.

Table 1 shows that the loss in efficiency remains very limited. When using the 0/1 weights, the efficiencies are above 90%, with the exception of

the setting where the sample size is low w.r.t. the dimension, i.e. $p = 5$ and $n = 50$. We also observe that the finite-sample efficiencies converge to their asymptotic counterparts. Using the smooth weight function $W_3$ yields slightly lower efficiencies, but they are more stable with respect to the sample size. We observe again convergence to the asymptotic efficiencies. We conclude from this simulation study that the loss in statistical efficiency when using the WML is very limited, while, as shown in the previous section, it has much better robustness properties.

## 8. Conclusion

To the best of our knowledge, this is the first paper where robustness for ordinal regression is studied. First we study the robustness of the classical Maximum Likelihood Estimator, and show that the slope estimator is explosion robust but implodes toward zero when well-chosen outliers are added to the data. We also showed that the ML estimator has an unbounded influence function, but the IF remains bounded with respect to contamination in the ordinal response variable. To obtain a bounded influence estimator, it therefore suffices to add weights based on the outlyingness of the values of the explanatory variables. The resulting Weighted Maximum Likelihood estimator has a bounded IF if the weight function is appropriately chosen. The price for the gain in robustness when using the WML is a loss in statistical efficiency. However, as shown in Section 7, this loss in efficiency remains limited.

The influence functions are not only useful in their own right but may also be used to compute asymptotic variances of the classical and robust

estimators, as was done in Section 7. Furthermore, the influence functions can be used in a diagnostic context, as shown in Section 6: combining robust distances with influence measures, one can detect different types of outliers (vertical outliers, good and bad leverage points) in a single diagnostic plot.

## Appendix

*Expression for the constant matrix $M(\psi, H_0)$ in* (12)

The matrix $M(\psi, H_0)$ can be decomposed as

$$\left(\begin{array}{cccccc|c} m_{11} & m_{12} & m_{13} & m_{14} & \dots & m_{1(J-1)} & M_{1B} \\ m_{12} & m_{22} & 2\gamma_2 m_{13} & 2\gamma_2 m_{14} & \dots & 2\gamma_2 m_{1(J-1)} & M_{2B} \\ m_{13} & 2\gamma_2 m_{13} & m_{33} & 2\gamma_3 m_{14} & \dots & 2\gamma_3 m_{1(J-1)} & M_{3B} \\ \vdots & \vdots & & \ddots & & \vdots & \vdots \\ m_{1(J-1)} & 2\gamma_2 m_{1(J-1)} & 2\gamma_3 m_{1(J-1)} & \dots & \dots & m_{(J-1)(J-1)} & M_{(J-1)B} \\ \hline M_{1B}^t & M_{2B}^t & \dots & \dots & \dots & M_{(J-1)B}^t & M_{BB} \end{array}\right).$$

Let $X \sim G_H$, then

$$m_{11} = -E_{G_H}\left[\sum_{j=1}^{J} \frac{(f(\alpha_j - \beta^t X) - f(\alpha_{j-1} - \beta^t X))^2}{F(\alpha_j - \beta^t X) - F(\alpha_{j-1} - \beta^t X)}\right],$$

$$m_{1k} = -2\gamma_k \left\{ E_{G_H}\left[\left(\frac{f(\alpha_k - \beta^t X)(f(\alpha_k - \beta^t X) - f(\alpha_{k-1} - \beta^t X))}{F(\alpha_k - \beta^t X) - F(\alpha_{k-1} - \beta^t X)}\right)\right] + \sum_{j=k+1}^{J} E_{G_H}\left[\left(\frac{(f(\alpha_j - \beta^t X) - f(\alpha_{j-1} - \beta^t X))^2}{F(\alpha_j - \beta^t X) - F(\alpha_{j-1} - \beta^t X)}\right)\right]\right\},$$

$$M_{1B} = E_{G_H}\left[X \sum_{j=1}^{J} \frac{(f(\alpha_j - \beta^t X) - f(\alpha_{j-1} - \beta^t X))^2}{F(\alpha_j - \beta^t X) - F(\alpha_{j-1} - \beta^t X)}\right],$$

$$M_{kB} = 2\gamma_k \left\{ E_{G_H}\left[X\left(\frac{f(\alpha_k - \beta^t X)(f(\alpha_k - \beta^t X) - f(\alpha_{k-1} - \beta^t X))}{F(\alpha_k - \beta^t X) - F(\alpha_{k-1} - \beta^t X)}\right)\right] + \sum_{j=k+1}^{J} E_{G_H}\left[\left(\frac{(f(\alpha_j - \beta^t X) - f(\alpha_{j-1} - \beta^t X))^2}{F(\alpha_j - \beta^t X) - F(\alpha_{j-1} - \beta^t X)}\right)\right]\right\},$$

for $k = 2, \ldots, J-1$. Furthermore, the diagonal elements are given by

$$m_{kk} = -4\gamma_k^2 \left\{ E_{G_H}\left[\left(\frac{f(\alpha_k - \beta^t X)^2}{F(\alpha_k - \beta^t X) - F(\alpha_{k-1} - \beta^t X)}\right)\right] + \sum_{j=k+1}^{J} E_{G_H}\left[\left(\frac{(f(\alpha_j - \beta^t X) - f(\alpha_{j-1} - \beta^t X))^2}{F(\alpha_j - \beta^t X) - F(\alpha_{j-1} - \beta^t X)}\right)\right]\right\}$$

for $k = 2, \ldots, J-1$, and

$$M_{BB} = -E_{G_H}\left[X^t X \sum_{j=1}^{J} \frac{(f(\alpha_j - \beta^t X) - f(\alpha_{j-1} - \beta^t X))^2}{F(\alpha_j - \beta^t X) - F(\alpha_{j-1} - \beta^t X)}\right].$$

To obtain the constant matrix in (13), we add the weight $W(D_H(X))$ in all the expectations.

*Proofs of the Propositions*

**Proof of Proposition 2** In order to prove this, let us show that for every finite number $m$ of outliers, there exists a real positive constant $M(Z_n, m)$

such that

$$\sup_{z_{n+1},\ldots,z_{n+m}} \|\hat{\theta}(Z'_{n+m})\| < M(Z_n, m).$$

For every $\theta = (\alpha_1, \ldots, \alpha_{J-1}, \beta^t)^t$, define

$$\delta(\theta, Z_n) = \inf\{\rho > 0 | \exists j \in \{1, \ldots, J-1\} \text{ and } \exists i \in I^{j+1} : x_i^t\beta \leq -\rho + \alpha_j \text{ or } i \in I^j : x_i^t\beta \geq \rho + \alpha_j\}.$$

The existence conditions stated in Proposition 1 imply that $0 < \delta(\theta, Z_n) < +\infty$. This can also be written as

$$\max_{j=1,\ldots,J-1} \{r_{j_0}(\theta), -r_{j_1}(\theta)\},$$

where $r_{j_0} = \min_{I^j} \max(x_i^t\beta - \alpha_j, 0)$ and $r_{j_1} = \max_{I^{j+1}} \min(x_i^t\beta - \alpha_j, 0)$. Thus, the mapping $\theta \to \delta(\theta, Z_n)$ is continuous. Then, with $\mathcal{S}^{p+J-2}$ denoting the sphere in $\mathbb{R}^{p+J-1}$ centered at the origin, one gets $\delta^*(Z_n) = \inf_{\theta \in \mathcal{S}^{p+J-2}} \delta(\theta, Z_n) > 0$.

Denote the log-likelihood function on the contaminated sample $Z'_{n+m}$ as $l$. Let $l_0$ be this log-likelihood computed for the vector $\theta^* = (\alpha^*, \beta^*)$ with $\alpha_j^* = F^{-1}(j/J), j = 1, \ldots, J-1$ and $\beta^* = 0$. It is easy to check that $l_0 = -(n+m)\log J$. Take $\tilde{z} = \exp(l_0)$ and define

$$M(Z_n, m) = \frac{F^{-1}(1-\tilde{z})}{\delta^*(Z_n)},$$

which only depends on the original sample $Z_n$ and on the number $m$ of outliers added to $Z_n$. Let us suppose that $\hat{\theta}_{n+m}$ satisfies

$$\|\hat{\theta}_{n+m}\| > M(Z_n, m). \tag{14}$$

For all $\hat{\theta}_{n+m} \in \mathbb{R}^{J-1+p}$, there exist at least one $j_0 \in \{1, \ldots, J-1\}$ and one $1 \leq i_0 \leq n$ such that

$$i_0 \in I^{j_0} \text{ and } \frac{x_{i_0}^t\hat{\beta}_{n+m} - \widehat{\alpha_{j_0}}_{n+m}}{\|\hat{\theta}_{n+m}\|} \geq \delta\left(\frac{\hat{\theta}_{n+m}}{\|\hat{\theta}_{n+m}\|}, Z_n\right) \geq \delta^*(Z_n) > 0 \tag{15}$$

or

$$i_0 \in I^{j_0+1} \text{ and } \frac{x_{i_0}^t \hat{\beta}_{n+m} - \widehat{\alpha_{j_0}}_{n+m}}{\|\hat{\theta}_{n+m}\|} \leq -\delta\left(\frac{\hat{\theta}_{n+m}}{\|\hat{\theta}_{n+m}\|}, Z_n\right) \leq -\delta^*(Z_n) < 0. \tag{16}$$

*Case* 1: When $i_0$ verifies (15), the log-likelihood is such that

$$l(\hat{\theta}_{n+m}; Z'_{n+m}) = \sum_{i=1}^{n+m} L(\hat{\theta}_{n+m}; z_i) \leq L(\hat{\theta}_{n+m}; z_{i_0})$$

since $L(\theta; z_i) = \sum_{j=1}^{J} \delta_{ij} \log\left(F(\alpha_j - \beta^t x_i) - F(\alpha_{j-1} - \beta^t x_i)\right)$ is always negative. Thus,

$$\begin{aligned} l(\hat{\theta}_{n+m}; Z'_{n+m}) &\leq \log\left[F\left(\widehat{\alpha_{j_0}}_{n+m} - x_{i_0}^t \hat{\beta}_{n+m}\right) - F\left(\widehat{\alpha_{j_0-1}}_{n+m} - x_{i_0}^t \hat{\beta}_{n+m}\right)\right] \\ &\leq \log\left[F\left(\widehat{\alpha_{j_0}}_{n+m} - x_{i_0}^t \hat{\beta}_{n+m}\right)\right] \leq \log\left[F\left(-\|\hat{\theta}_{n+m}\|\delta^*(Z_n)\right)\right] \\ &= \log\left[1 - F\left(\|\hat{\theta}_{n+m}\|\delta^*(Z_n)\right)\right] < \log\left[1 - F\left(M(Z_n, m)\delta^*(Z_n)\right)\right] \\ &= \log(\tilde{z}) = l_0 \end{aligned}$$

using the inequalities in (15), the symmetry and strict increasing behavior of $F$, the hypothesis that $\|\hat{\theta}_{n+m}\| > M(Z_n, m)$ and the definitions of $M(Z_n, m)$ and $\tilde{z}$.

*Case* 2: When $i_0$ verifies (16), one also gets $l(\hat{\theta}_{n+m}; Z'_{n+m}) < l_0$. Here the inequality $\log(y - x) < \log(1 - x)$ for $0 < x < y < 1$ is used.

The inequality $l(\hat{\theta}_{n+m}; Z'_{n+m}) < l_0$ implies that $\hat{\theta}_{n+m}$ cannot be the ML-estimate. Therefore, equation (14) does not hold and the theorem is proven.

**Proof of Proposition 3** As in the previous proof, let $l$ be the log-likelihood function on the contaminated sample $Z'_{n+m}$ and $l_0$ its value computed for $\theta^*$.

Let $\delta > 0$ be fixed. It is always possible to find $\xi > 0$ s.t. $\log(F(-\xi)) = l_0$. Let us define $M = \max_{1 \leq i \leq n} \|x_i\|$, $N = \xi/\delta$ and $A = \sqrt{p}(2N + M)$. Take $\{e_1, \ldots, e_p\}$ the canonical basis of $\mathbb{R}^p$ and add to the initial sample $Z_n$ the $m = pJ$ outliers

$$z_i^j = (v_i^t, j) \text{ with } v_i = A\, e_i,\ i = 1, \ldots, p \text{ and } j = 1, \ldots, J.$$

Take $\|\beta\| > \delta$, $\alpha$ arbitrarily and $\theta = (\alpha, \beta)$. The aim is to show that $l(\theta; Z'_{n+m}) < l_0$ as soon as $\|\beta\| > \delta$ which will imply that $\|\hat{\beta}_{n+m}\| \leq \delta$. Since this reasoning holds for all $\delta > 0$, $\hat{\beta}$ can be made as small as possible by adding $m = pJ$ outliers.

For $j$ fixed in $\{1, \ldots, J-1\}$, define the hyperplane $H_\delta^j = \{x \in \mathbb{R}^p : \alpha_j - x^t\beta = 0\}$. The distance between a vector $x \in \mathbb{R}^p$ and $H_\delta^j$ is

$$\text{dist}(x, H_\delta) = \left| x^t \frac{\beta}{\|\beta\|} - \frac{\alpha_j}{\|\beta\|} \right|.$$

First, suppose that $\exists\, i_0 \in \{1, \ldots, p\}$ s.t. $\text{dist}(v_{i_0}, H_\delta^j)$ is bigger than $N$. If $\alpha_j - v_{i_0}^t\beta > 0$, then take the outlier $z_{i_0}^{j+1}$. Since $\alpha_j - v_{i_0}^t\beta = \text{dist}\left(v_{i_0}, H_\delta^j\right) \|\beta\| > N\|\beta\| > N\delta = \xi$, one has

$$\begin{aligned} l(\theta, Z'_{n+m}) &\leq L(\theta, z_{i_0}^{j+1}) = \log\left[F\left(\alpha_{j+1} - v_{i_0}^t\beta\right) - F\left(\alpha_j - v_{i_0}^t\beta\right)\right] \\ &\leq \log\left[1 - F\left(\alpha_j - v_{i_0}^t\beta\right)\right] < \log(1 - F(\xi)) = \log(F(-\xi)) = l_0. \end{aligned}$$

On the other hand, if $\alpha_j - v_{i_0}^t\beta < 0$, then take the outlier $z_{i_0}^j$. Since $-\left(\alpha_j - v_{i_0}^t\beta\right) > N\|\beta\| > N\delta = \xi$, one has

$$\begin{aligned} l(\theta, Z'_{n+m}) &\leq L(\theta, z_{i_0}^{j}) = \log\left[F\left(\alpha_j - v_{i_0}^t\beta\right) - F\left(\alpha_{j-1} - v_{i_0}^t\beta\right)\right] \\ &\leq \log\left[F\left(\alpha_j - v_{i_0}^t\beta\right)\right] < \log(F(-\xi)) = l_0. \end{aligned}$$

Now, suppose that the distance between the hyperplane $H_\delta^j$ and $v_k$ is smaller than $N$ for all $k = 1, \ldots, p$. Let $k_0$ be the index s.t. $|\beta_{k_0}| = \max_{1 \leq k \leq p} |\beta_k|$. If $\beta_{k_0} > 0$, it follows that $\beta_{k_0} A - \alpha_j = \text{dist}(v_{k_0}, H_\delta^j)\|\beta\| \leq N\|\beta\|$ and $\beta_{k_0} \geq \|\beta\|/\sqrt{p}$ leading to

$$\alpha_j \geq \|\beta\| \left( \frac{A}{\sqrt{p}} - N \right) = (M + N)\|\beta\|.$$

Therefore, take an observation $z_{i_0}$ from $Z_n$ with $y_{i_0} = j + 1$. Now,

$$\alpha_j - \beta^t x_{i_0} \geq \alpha_j - \|\beta\|\|x_{i_0}\| \geq (M + N)\|\beta\| - M\|\beta\| > N\delta = \xi$$

and

$$\begin{aligned} l(\theta, Z'_{n+m}) \leq L(\theta, z_{i_0}) &= \log \left[ F\left(\alpha_{j+1} - \beta^t x_{i_0}\right) - F\left(\alpha_j - \beta^t x_{i_0}\right) \right] \\ &\leq \log \left[ 1 - F\left(\alpha_j - \beta^t x_{i_0}\right) \right] < \log(1 - F(\xi)) = \log(F(-\xi)) = l_0 \end{aligned}$$

On the other hand, if $\beta_{k_0} < 0$, then $\alpha_j - \beta_{k_0} A \leq |\alpha_j - \beta_{k_0} A| \leq N\|\beta\|$ and $\beta_{k_0} \leq -\|\beta\|/\sqrt{p}$ which leads to

$$\alpha_j \leq \|\beta\| \left( N - \frac{A}{\sqrt{p}} \right) = -(M + N)\|\beta\|.$$

Therefore, take an observation $z_{i_0}$ from $Z_n$ with $y_{i_0} = j$. Now,

$$\alpha_j - \beta^t x_{i_0} \leq -(N + M)\|\beta\| - \beta^t x_{i_0} \leq -(M + N)\|\beta\| + M\|\beta\| < -N\delta = -\xi$$

and

$$\begin{aligned} l(\theta, Z'_{n+m}) \leq L(\theta, z_{i_0}) &= \log \left[ F\left(\alpha_j - \beta^t x_{i_0}\right) - F\left(\alpha_{j-1} - \beta^t x_{i_0}\right) \right] \\ &\leq \log \left[ F\left(\alpha_j - \beta^t x_{i_0}\right) \right] < \log(F(-\xi)) = l_0. \end{aligned}$$

As this reasoning holds for all $j \in \{1, \ldots, J - 1\}$, the theorem is proven.

## References


Albert, A., Anderson, J.A., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1–10.

Anderson, J.A., Philips, P.R., 1981. Regression, discrimination and measurement models for ordered categorical variables. J. Roy. Statist. Soc. Ser. C 30, 22–31.

Bastien, P., Esposito Vinzi, V., Tenenhaus, M., 2005. PLS generalised linear regression. Comput. Statist. Data Anal. 48, 17–46.

Bianco, A.M., Yohai, V.J., 1996. Robust estimation in the logistic regression model, in: Robust statistics, data analysis, and computer intensive methods (Schloss Thurnau, 1994). Springer, New York. volume 109 of *Lecture Notes in Statist.*, pp. 17–34.

Bondell, H.D., 2008. A characteristic function approach to the biased sampling model, with application to robust logistic regression. J. Statist. Plann. Inference 138, 742–755.

Carroll, R.J., Pederson, S., 1993. On robustness in the logistic regression model. J. Roy. Statist. Soc. Ser. B 55, 693–706.

Croux, C., Flandre, C., Haesbroeck, G., 2002. The breakdown behavior of the maximum likelihood estimator in the logistic regression model. Statist. Probab. Lett. 60, 377–386.

Croux, C., Haesbroeck, G., 2003. Implementing the Bianco and Yohai estimator for logistic regression. Comput. Statist. Data Anal. 44, 273–295.

Croux, C., Haesbroeck, G., Joossens, K., 2008. Logistic discrimination using robust estimators: an influence function approach. Canad. J. Statist. 36, 157–174.

Franses, P.H., Paap, R., 2001. Quantitative Models in Marketing Research. Cambridge University Press.

Gervini, D., 2005. Robust adaptive estimators for binary regression models. J. Statist. Plann. Inference 131, 297–311.

Haberman, S.J., 1980. Discussion of McCullagh (1980). J. Roy. Statist. Soc. Ser. B 42, 136–137.

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

Hobza, T., Pardo, L., Vajda, I., 2008. Robust median estimator in logistic regression. J. Statist. Plann. Inference 138, 3822–3840.

Hosseinian, S., Morgenthaler, S., 2011. Robust binary regression. J. Statist. Plann. Inference 141, 1497–1509.

Müller, C.H., Neykov, N., 2003. Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. J. Statist. Plann. Inference 116, 503–519.

Pison, G., Van Aelst, S., 2004. Diagnostic plots for robust multivariate methods. J. Comput. Graph. Statist. 13, 310–329.

Powers, D.A., Xie, Y., 2008. Statistical methods for categorical data analysis. Emerald.

Rousseeuw, P., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust regression and outlier detection. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

Wang, C.Y., Carroll, R.J., 1995. On robust logistic case-control studies with response-dependent weights. J. Statist. Plann. Inference 43, 331–340.