Robust Inference in Generalized Linear Models

FATEMAH ALQALLAF¹ AND CLAUDIO AGOSTINELLI²

¹Department of Statistics and Operations Research, Kuwait University, Safat, Kuwait
²Department of Mathematics, University of Trento, Trento, TN, Italy

Robust inference on the parameters in generalized linear models is performed using the weighted likelihood method. Two cases are considered: a case with replicated observations and a case with a single observation of the dependent variable for each combination of the explanatory variables. The first case is common in the design of experiments, while the second case arises in observational studies. Theoretical and computational results on real datasets are presented and compared with other existing techniques.

Keywords Gamma model; Inverse Gaussian model; Outliers; Poisson model; Robust estimation; Robust model selection; Weighted likelihood.

Mathematics Subject Classification Primary 62F35; 62FG35; Secondary 62-07.

1. Introduction

Generalized linear models (GLMs) are commonly used in many different fields and inferences are made by using the likelihood function and related quantities. The likelihood method has good asymptotic properties when the true model generating the data is the same as the hypothesized model. However, this approach leads to unreliable results when the hypothesized model does not coincide with the true model. Robust procedures are proposed by several authors in order to obtain more stable inferences. The M-estimator procedure for binary response variable was first introduced by Pregibon (1982), Stefanski et al. (1986), and Copas (1988) (see also Carroll and Pederson, 1993; Künsch et al., 1989). A similar inference approach was used by Bianco and Yohai (1996) in order to have a consistent and more robust estimator. Croux and Haesbroeck (2003) further improved and presented a fast implementation scheme. M-estimators based on quasi likelihood were introduced by Cantoni and Ronchetti (2001). Croux et al. (2002) studied the behavior of maximum likelihood estimators in presence of outliers, while Rousseeuw and Christmann (2003) provided an approach for the logistic regression model. Markatou et al. (1997) introduced...
the weighted likelihood approach for discrete models with particular attention to the logistic regression case, while Markatou et al. (1998) extended the method for continuous models. Hosseinian (2009) studied a different form of weighted likelihood and applied it to binary and Poisson regression. Ruckstuhl and Welsh (2001) compared the minimum distance methods with other robust methods. Muller and Neykov (2003) used the trimmed likelihood for generalized linear models. Most of these methods are designed for logistic regression and Poisson regression and cannot be easily extended to other generalized linear models.

In this article, we investigate the use of weighted likelihood method to obtain robust estimation, inference, and model selection for generalized linear models with distributions in the regular exponential family. As in Markatou et al. (1998) outliers are “observations that are highly unlikely to occur under the hypothesized model.” This concept is different from the usual “geometric” distance-based outlier because it depends on a “probabilistic” measure.

In the context of regression, the above type of outliers is often called vertical outliers since they affect the response variable, while horizontal outliers affect the explanatory variables and produce leverage points. We show through Monte Carlo experiments that the proposed procedure is robust against vertical outliers. The effect of horizontal outliers is not investigated here.

Even though there exist some studies on the weighted likelihood method for replicated observations, to the best of our knowledge, there is no prior literature on weighted likelihood methods in GLM for few or no replicated observations. The new method utilizes Anscombe residuals and is implemented in the R package \texttt{wle} available on CRAN.

The article is organized as follows. Section 2 discusses the estimation procedure, with particular emphasis on the use of Anscombe residuals. Section 3 is devoted to inference, analysis of variance decomposition, and model selection. Section 4 provides an example based on a real dataset. Some computational results obtained from Monte Carlo simulations for Poisson, gamma, and inverse Gaussian models are reported in Section 5. Finally, Section 6 contains conclusions and remarks. Supplemental material is available at www.dst.unive.it/~claudio/uploads/LSSP_2013_0307.

2. Robust Estimation in Generalized Linear Models

The weighted likelihood estimates (WLE; Markatou et al., 1998) are obtained as solutions of weighted likelihood estimating equations (WLEE) that are a modified type of the maximum likelihood estimating equations in which each score is associated with a weight that could depend on the unknown parameters. Lindsay (1994) shows that, for discrete models and certain types of distance/divergence measures between distributions, the corresponding estimating equations could be written as WLEE with a particular form of the weight function. Markatou et al. (1998) extended this approach to continuous models. Hereafter, we review the construction of the weights for a random sample \( y_1, \ldots, y_n \) with a common unknown parameter \( \theta \). Later, we consider the extension to the generalized linear model, which has a sample of independent vectors \((y_1, x_1), \ldots, (y_n, x_n)\), where each response \( y_i \) has an expectation that depends on explanatory variable \( x_i \).

The weights are defined as follows:

\[
w(y; \theta, f^*) = \max \left( \min \left( \frac{A(\delta(y; \theta, f^*)) + 1}{\delta(y; \theta, f^*) + 1}, 1 \right), 0 \right),
\]
where $\delta(\cdot)$ is the Pearson residual function and $A(\delta)$ is the Residual Adjustment Function (RAF; Lindsay, 1994). Hence, the WLE is the solution of

$$
\sum_{i=1}^{n} w(y_i; \theta, f^*) u(y_i; \theta) = 0,
$$

where $u(y; \theta)$ is the classical score function from the hypothesized model. Whenever possible we use the simpler notation $w_i(\theta)$ for $w(y_i; \theta, f^*)$.

The Pearson residual function $\delta(y, \theta, f^*) = f^*(y) m^*(y; \theta) - 1$ shows the agreement between the hypothesized model and the empirical distribution of the observations, where $f^*(y) = \int k(y; t, h) d\hat{F}_n(y)$ is a nonparametric density estimate based on the data, $k(y; t, h)$ is a kernel with bandwidth $h$, and $\hat{F}_n(y)$ is the empirical distribution function given from the sample. $m^*(y; \theta) = \int k(y; t, h) dM(y; \theta)$ is a smoothed version of the density model for a continuous random variable, whereas, for a discrete model the smoothing is not needed. A discussion on the bandwidth selection is presented in the Appendix.

The residual adjustment function operates on Pearson residuals as the Huber $\psi$-function operates on the structural residuals. When $A(\delta) = \delta$ then the weight $w_i(\theta) \equiv 1$, and the weighted likelihood estimating equations correspond to the maximum likelihood estimating equations. The literature contains several proposals for the RAF, here we recall two different families. The RAF based on the Power Divergence Measure (PDM; Cressie and Read, 1984, 1988) is

$$
A_{\text{PDM}}(\delta, \tau) = \begin{cases} 
\tau (\delta + 1)^{1/\tau} - 1 & \tau < \infty \\
\log(\delta + 1) & \tau \to \infty
\end{cases}
$$

and was introduced by Lindsay (1994). Special cases are maximum likelihood ($\tau = 1$), Hellinger distance ($\tau = 2$), Kullback–Leibler divergence ($\tau \to \infty$), and Neyman’s Chi-Square ($\tau = -1$). The RAF based on the Generalized Kullback–Leibler divergence (GKL) was introduced in Park and Basu (2003)

$$
A_{\text{GKL}}(\delta, \tau) = \frac{\log(\tau \delta + 1)}{\tau} 
$$

$0 \leq \tau \leq 1 \, .$

Special cases are maximum likelihood ($\tau \to 0$) and Kullback–Leibler divergence ($\tau = 1$). This RAF can be interpreted as a linear combination between the Likelihood divergence and the Kullback–Leibler divergence. We will use this last family in all our subsequent examples.

As an illustration, Fig. 1 displays the Pearson residual function evaluated at $\theta = (24, 8)$ for the gamma family of shape $s$ and rate $r$ with density $m(y; s, r)$ (see central panel), the contaminated distribution $f(y) = 0.95 \, m(y; 24, 8) + 0.05 \, m(y; 48, 32)$ (see left panel), and the weight function using the generalized Kullback-Leibler RAF with parameter $\tau = 0.25$ (see right panel). The figure shows that the interval with high probability for the contamination distribution (dotted line) is well identified by the Pearson residual function and the observations in that interval are heavily down-weighted.

**Example 2.1 (Movimprese dataset).** The weighted likelihood estimate ($\hat{\lambda}_{w}$) for the parameter $\lambda$ of a Poisson distribution is the solution of the following fixed point equation

$$
\lambda = \frac{\sum_{i=1}^{n} w_i(\lambda) y_i}{\sum_{i=1}^{n} w_i(\lambda)} .
$$
on the linear predictor $\eta$ is the $x$ $\beta$ does not depend on $\phi$.

Figure 1. Left panel: $f(y) = 0.95 m(y; 24, 8) + 0.05 m(y; 48, 32)$, where $m(y; s, r)$ is the density of a gamma distribution with shape $s$ and rate $r$. Center panel: Pearson residual function ($\delta(y; 24, 8, f(y))$). Right panel: Weight function ($w(y; 24, 8, f(y))$).

We illustrate the method of WLE using a dataset from the Movimprese database available at http://www.infocamere.it/movimprese.htm. We consider the number of new individual companies (SPISC) in the Chambers of Commerce together with the total registered individual companies (SPREG) in the agriculture sector for each trimester (TRIM), from 1995 to 2009 for the four provinces (PROV) of the Friuli Venezia Giulia region: Gorizia (GO), Pordenone (PN), Trieste (TS), and Udine (UD). The response variable is the SPISC that can be fit with the Poisson distribution. Here, we use only the observations related to the Udine province with 60 observations. The WLE using GKL RAF with $\tau = 0.1$ is 5.097 and the maximum likelihood estimate is 13.733. The left panel of Fig. 2 shows the conditional distribution of SPISC for the Udine province, while the right panel displays the final weights. It is evident from the figure that there are unusual observations in the right tail of the distribution, which is incompatible with the Poisson model for the bulk of the data, and these observations are heavily down-weighted by the WLE procedure. Notice that removing the observations with weight smaller than 0.1, which could be considered as outliers, reduces the MLE to 5.347.

In a generalized linear model (McCullagh and Nelder, 1989), a response $y$ has density $m(y; \theta) \propto \exp((y\theta - b(\theta))/\phi + c(y, \phi))$, where $\theta$ is the natural parameter and $\phi$ is the dispersion parameter. Let $\mu = E(y) = b'(\theta)$ be the mean function and $\text{Var}(y) = V(\mu)\phi$ the variance of $y$, where $V(\mu) = b''(\theta)$ is the variance function. The mean function $\mu$ depends on the linear predictor $\eta = x^t\beta$ through the monotone link function $g$ with $\eta = g(\mu)$. Here $x$ is the $p$-vector of the explanatory variables and $\beta$ is a $p$-vector of the parameters.

Given $n$ independent responses $y = (y_1, y_2, \ldots, y_n)$ and the corresponding design matrix $X$ with dimension $n \times p$, the log likelihood is $\ell(\beta) = \sum_{i=1}^n (y_i\theta_i - b(\theta_i))/\phi_i + c(y_i, \phi_i)$, where $\theta_i = \theta(\eta_i)$ with $\eta_i = x_i^t\beta$. By differentiation, we obtain the score function $\partial\ell(\beta)/\partial \beta = \partial \eta^t/\partial \beta \partial \theta/\partial \eta^t \partial \theta/\partial \beta^t \ell(\theta) = X'\mu(\beta)$, where the design matrix $X = \partial \eta/\partial \beta'$ does not depend on $\beta$. Since the score equation $X'\mu(\beta) = 0$ is nonlinear, an iterative re-weighted least-square algorithm (IRLS) is often used with the adjusted dependent variable $z = X\beta + g'(\mu)(y - \mu)$ regressed on the columns of $X$ using the weight diagonal matrix $\Omega = \Omega(\beta)$ with diagonal elements $\omega_i(\beta) = [g'(\mu_i)^2\phi_i V(\mu_i)]^{-1}$. The maximum likelihood estimates are obtained by repeating the IRLS step until convergence is achieved.

The weighted likelihood estimating equations are obtained by attaching a weight diagonal matrix $W(\beta)$, with elements $w_i(\beta)$ based on the Pearson residuals, to each score contribution, that is, the weighted likelihood estimates are the solutions of the equation $X'W(\beta)\mu(\beta) = 0$. The solutions can be obtained by a two-step algorithm. The first step...
evaluates the matrix $W(\beta)$ and the second step is the IRLS algorithm with $W = W(\beta)$ kept fixed, as follows:

$$\beta = (X'W^{1/2}\Omega W^{1/2}X)^{-1}X'W^{1/2}\Omega(W^{1/2}X\beta + \Omega^{-1}W^{1/2}u(\beta))$$

and, since both $W$ and $\Omega$ are diagonal matrix, a simple expression for $\beta$ is

$$\beta = (X'\Omega WX)^{-1}X'\Omega W(X\beta + \Omega^{-1}u(\beta)) = (X'\Omega WX)^{-1}X'\Omega Wz.$$ 

Thus, the weight matrix $\Omega(\beta)$ used in the IRLS algorithm for the MLE is replaced by a new weight diagonal matrix with elements $\omega_j(\beta)w_j(\beta)$. The IRLS algorithm is iterated until convergence is attained, then the matrix $W(\beta)$ is updated and is inserted in the IRLS. The two-step process terminates when the values of $W(\beta)$ elements remain unchanged.

To highlight the definition of the weights hereafter, we use a different notation. Let $Y_j \sim m(\cdot, \theta_j(\beta), \phi_j)$, $j = 1, \ldots, J$ be $J$ independent random variables and $y_j = (y_{j,1}, y_{j,2}, \ldots, y_{j,n_j})$ be $n_j$ replicates from the random variable $Y_j$. Hence $y$ equals to $(y_1, y_2, \ldots, y_J)$. The weight $w_{jk}(\beta) = w(y_{jk}, \theta_j(\beta), \phi_j, f^{*}_{j}(\cdot))$ is based on the Pearson residual function evaluated at the conditional distribution of $Y_j$, that is,

$$\delta(y_{jk}; \theta_j(\beta), \phi_j) = \frac{f^{*}_{j}(y_{jk})}{m^{*}(y_{jk}; \theta_j(\beta), \phi_j)} - 1,$$

where $f^{*}_{j}(\cdot)$ is a nonparametric estimate of the conditional density (probability distribution in the case of discrete models) based on the $n_j$ replicates in the $j$ level. If $X$ is a fixed (balanced) design matrix with a small number of distinct values $J$ of the parameters ($\theta, \phi$) compared to the sample size $n$, then we will have an adequate number of observations $n_j$ to estimate all the distinct $f^{*}_{j}(\cdot)$ conditional distributions. In practice, problems arise when the sample size $n$ is small with respect to $J$ and this implies small values of $n_j$ for all $j$ in the balanced design and for the unbalanced design some $n_j$ might be very small. In either cases “good” nonparametric estimates $f^{*}_{j}(\cdot)$ for some or all conditional distributions may not be available. A naive solution is to use observations that correspond to ($\theta, \phi$) values in the neighborhood of ($\theta_j, \phi_j$) in order to estimate $f^{*}_{j}(\cdot)$ at $y_{jk}$. We adopt a more general solution using the Anscombe residuals $r_{jk}$ ($j = 1, \ldots, J, k = 1, \ldots, n_j$) because their asymptotic distribution, under the hypothesized model, is as normal as possible. The Anscombe residuals were introduced in Cox and Snell (1968) and Loynes (1969) (see also Brant, 1987; Pierce and Schafer, 1986) and obtained by the transformation $R'(y) = \int_{-\infty}^{y} V(\mu)^{-1/3} d\mu$ in both $y$ and $\mu$, adjusted for the scale $R'(\mu)N(\mu)^{1/2}$, that is,

$$A(y, \mu) = \frac{R(y) - R(\mu)}{R'(\mu)N(\mu)^{1/2}}.$$ 

Anscombe residuals were also considered in McCullagh and Nelder (1989, sec. 2.4.2). Using the Anscombe residuals allows us to estimate only one nonparametric distribution, which will be common to all residuals regardless of the $j$ values. Let $r_{jk} = A(y_{jk}; \mu_j(\beta), \phi_j)$ be the Anscombe residuals and $\hat{F}_n(r)$ be the corresponding empirical distribution. We let $\hat{f}(r) = \int k(r; t, h) d\hat{F}_n(t)$ be the nonparametric density estimate. An estimate of the smoothed model $\hat{m}(\cdot)$ of the Anscombe residuals distribution is obtained from simulation and is described as follows. Let $\hat{p}_j = n_j/n$ be the proportion of observations in each $j$ level. In each iteration of the two-step algorithm, we sample $v$ observations ($z_1, \ldots, z_v$) from
the mixture model \( \sum_{j=1}^{J} \hat{p}_j \ m(z, \hat{\theta}_j(\hat{\beta}), \hat{\phi}_j) \) and we calculate their Anscombe residuals \((r_1, \ldots, r_v)\) with the empirical distribution function \( \tilde{F}_v(\cdot) \). Then we have a consistent estimate of \( \tilde{m}(\cdot) \) given by \( \tilde{m}(r) = \int k(r; t, h) \ d\tilde{F}_v(t) \). Comparing \( \tilde{f}(r) \) with \( \tilde{m}(r) \) provides the Pearson residuals

\[
\delta_A(r_{jk}; \hat{\theta}_j(\hat{\beta}), \hat{\phi}_j) = \frac{\tilde{f}(r_{jk})}{\tilde{m}(r_{jk}; \hat{\theta}_j(\hat{\beta}), \hat{\phi}_j)} - 1.
\]

However, the calculations can be simplified often since in most situations, the distribution \( \tilde{m}(\cdot) \) of the Anscombe residuals is well approximated by the standard normal distribution. See the supplementary material and Section 4 for an empirical evidence about this fact.

In most generalized linear models, the \( \phi_i \) are unknown, for some cases \( \phi_i = a_i \phi \) with known constants \( a_i \). For unknown \( \phi \), a robust estimate can be obtained using a weighted likelihood estimating equation, with weight matrix \( W(\beta) \) as defined above and score function determined by the model. Alternatively, a robust estimate of the variance \( \text{Var}(y) \) can be defined leading to the following robust method of moments estimates

\[
\hat{\phi} = \frac{1}{\sum_{i=1}^{n} w_i(\hat{\beta})} \left( \sum_{i=1}^{n} \frac{w_i(\hat{\beta})}{a_i V(\hat{\mu}_i)} (y_i - \hat{\mu}_i)^2 \right),
\]

where \( \hat{\beta} \) and \( \hat{\mu}_i \) are the WLE estimates of the corresponding quantities. In the gamma family, practical experience suggests that the estimate based on weighted likelihood is more stable than the weighted moment estimate (1), while it is the opposite for the inverse Gaussian family. In most cases, the score function for \( \beta \) does not depend on \( \phi \), however, the weight matrix \( W(\beta) \) does depend on \( \phi \), hence an iterated algorithm must be used.

Starting values for the algorithm can be obtained by subsampling techniques. A subsample of small size is drawn from the original dataset and maximum likelihood estimates are obtained and used as the initial values in the weights for the original dataset. The subsampling is repeated several times searching for all possible roots of the weighted likelihood estimating equations. To find subsamples without outliers, the size of subsamples should be as small as possible, considering the number of the parameters that need to be estimated and algorithm stability. In practice, 20 subsamples with size equal the number of parameters plus 2 provide to be effective.

Anscombe residuals can be avoided in discrete models when all \( n_j \) are sufficiently large. In this case, the roots of the WLEE are equal to the stationary points of the related minimum distance/disparity approach as shown in Markatou et al. (1997). Hence, the appropriate solution can be identified as the root that minimizes the distance/disparity. For all other situations, the relation with the minimum distance/disparity approach is lost. Thus, the method in Agostinelli (2006) can be applied to choose the appropriate root.

Asymptotics and robust properties of the WLEE are studied in Markatou et al. (1997, 1998). Under the assumptions that the sample size for each distinct value of \((\theta, \phi)\) tends to infinity \((n_j \to \infty \text{ for all } j)\), and hence we do not use the Anscombe residuals, the asymptotic and robust properties for GLM in the exponential family can be obtained directly from Markatou et al. (1997, 1998). In particular, in the absence of contamination the estimators \( \hat{\beta} \) are consistent and asymptotically normally distributed with variance matrix \( (X'\Omega(\hat{\beta})X)^{-1} \), which leads to a first-order efficient estimator. Under contamination the distribution is not theoretically derived, but Monte Carlo simulations show that the distribution of \( \hat{\beta} \) can be well approximated by a normal distribution with variance matrix estimated as \( (X'\Omega(\hat{\beta})W(\hat{\beta})X)^{-1} \). We refer to Markatou et al. (1997, 1998) for the theoretical
details. If the use of the Anscombe residuals is unavoidable, that is, at least some of the \( n_j \) does not go to infinity, or the number of distinct value of \((\theta, \phi)\) tends to infinity with the same speed as the sample size, then asymptotic and robust properties have still to be established. In the Appendix, Theorem A.1, we prove that the Pearson Residuals based on Anscombe residuals converges to zero uniformly almost surely under the assumed model. This is sufficient to establish consistency and asymptotic normality for the estimators following the results in Markatou et al. (1997, 1998). Robust properties are studied through Monte Carlo simulation, see Section 5 and the supplemental material for complete results.

3. Robust Inference in Generalized Linear Models

Agostinelli and Markatou (2001) introduced robust alternatives of the classical tests: Wald test, Wilks likelihood ratio test, and Rao score test. Under no contamination, the robust tests are asymptotically equivalent to the corresponding classical tests. In the presence of contamination, level and power breakdown points (He et al., 1990) of the robust tests are equal to the breakdown point of the corresponding weighted likelihood point estimators. These tests can be extended to a generalized linear model framework. Given the following system of hypotheses

\[ H_0 : \beta \in B_0 \text{ vs. } H_1 : \beta \in B_1 \subseteq B/B_0 \]

(\( B \) is the parametric space), the weighted Wald-test is given by

\[ T_w = (\hat{\beta}_0 - \hat{\beta})' (X' \Omega (\hat{\beta}) W (\hat{\beta}) X)^{-1} (\hat{\beta}_0 - \hat{\beta}) \]

where \( \hat{\beta}_0 \) is the minimizer of \( \sum_{i=1}^{n} w_i(\hat{\beta}) \ell(y_i; \hat{\theta}(\beta), \phi_i) \) inside the parametric space restricted to the null hypothesis (\( \hat{\beta}_0 \in B_0 \)). Notice that the weights \( w_i(\hat{\beta}) \) are evaluated at the estimates of \( \beta \) in the full parametric space \( B \). It is easy to show (Agostinelli and Markatou, 2001) that the weights based on the restricted space \( B_0 \) can lead to inconsistent tests even though there is no contamination. Under the null hypothesis, the \( T_w \) test is approximately distributed as a \( \chi^2_{p-q} \), where \( q \) is the dimension of the space \( B_0 \).

Example 3.1 (Movimprese dataset, SPISC~PROV-1). A Poisson regression with log link of the variable SPISC on PROV (provinces) is estimated using MLE, WLE, and WLE\(_A\) (the weighted likelihood estimates with the Anscombe residuals). Table 1 shows the estimates of the coefficients, standard error (S.E.), and \( p \)-values (P-val.). The inference of WLE is made using a robust Wald test. The results of WLE and WLE\(_A\) are equivalent. We notice that there is a one point difference between MLE and WLE for the coefficients of Pordenone (PN) and Udine (UD), and approximately 1/3 in ratio for Gorizia (GO). Standard errors are always slightly larger in WLE than in MLE. Figure 3 displays the conditional distributions of SPISC against PROV and the conditional fitted distribution by MLE and WLE. The differences in the estimated coefficients lead to a poor fit for the provinces Pordenone (PN) and Udine (UD).

The weighted log-likelihood test for the above system of hypotheses is given by

\[ T_\ell = 2 \sum_{i=1}^{n} w_i(\hat{\beta}) \ell(y_i, \hat{\beta}) - \ell(y_i, \hat{\beta}_0) \]

which has an approximated \( \chi^2_{p-q} \) distribution under the null hypothesis, and is based on a weighted form of deviance. Following Agostinelli (2002), we define a weighted Akaike Information Criterion to perform robust model selection between nested models \( \text{WAIC}_\alpha = -2 \sum_{i=1}^{n} w_i(\hat{\beta}) \ell(y_i, \hat{\beta}_0) + \alpha q \), where \( \hat{\beta}_0 \) is defined as above, and \( p - q \) coefficients are treated as zeros. The constant \( \alpha \) is equal to 2 while if we put \( \alpha = \ln(n) \), we get a weighted Bayesian Information Criterion. It is necessary to use weights evaluated at the full model (\( \hat{\beta} \in B \)) to ensure asymptotic consistency of the method and to preserve the relation between the classical log-likelihood test and the Akaike Information Criterion in their weighted forms.
Let \( A \) and \( B \) denote two nested models included in the full model. A weighted deviance decomposition is defined as follows. Let \( \tilde{\eta} \) be the estimate of \( \eta \), under the saturated model, which is a model without constrains on the \( \eta \)'s. When \( \beta \) has dimension \( n \) and if the map between \( \beta \) and \( \eta \) is one-to-one, the maximization of the weighted likelihood (using weights fixed at the full model) does not depend on the weights. Then the scaled deviance for model \( A \) is defined as

\[
D(\hat{\beta}_A, \hat{\beta}) = 2 \sum_{i=1}^{n} w_i(\hat{\beta}) \left( \ell(y_i; \tilde{\eta}_i, \tilde{\phi}_i) - \ell(y_i; \hat{\eta}_A, \hat{\phi}_A) \right),
\]

where \( \hat{\beta}_A \) and the corresponding \( \hat{\eta}_A, \hat{\phi}_A \) are obtained by minimizing the weighted likelihood, with fixed weights at the full model \( \sum_{i=1}^{n} w_i(\hat{\beta}) \ell(y_i; \theta_i(\beta), \phi_i) \). A deviance table can be constructed from the deviance differences

\[
D(\hat{\beta}_B, \hat{\beta}) - D(\hat{\beta}_A, \hat{\beta}) = 2 \sum_{i=1}^{n} w_i(\hat{\beta}) \left( \ell(y_i; \hat{\eta}_A, \hat{\phi}_A) - \ell(y_i; \hat{\eta}_B, \hat{\phi}_B) \right),
\]

which has an approximated \( \chi^2 \) distribution with degrees of freedom equal to the difference of degrees of freedom between the two models \( B \) and \( A \). When \( \phi \) is known, a goodness-of-fit measure is the weighted Pearson’s statistic

\[
P_w = \frac{1}{\phi} \sum_{i=1}^{n} w_i(\hat{\beta}) \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/a_i},
\]

where \( \hat{\mu}_i \) are the mean fitted values.

4. Example: Number of New Enterprises in Friuli Venezia Giulia Region

We illustrate our robust procedures using the dataset in Example 2.1. Since 1982, the Info
camere (an IT Consortium of the Italian Chambers of Commerce, http://www.infocamere.
itwww.infocamere.it) has been providing on a regular bases information (Movimprese database) about the demography of Italian companies. For every quarter of the year, they
Figure 3. Movimprese dataset. Empirical conditional probability functions (solid), MLE-fitted values (dotted), WLE-fitted values (dashed) using the model $SPISC \sim PROV - 1$. The plots are right truncated to display the main part of the distribution.

report the birth/death count of the registered companies in the Chambers of Commerce stratified by economic sector, legal form, and region. We consider the data of the agricultural sector and the Friuli Venezia Region during the period of 1995–2009. In 1995, the database was transferred to a new platform and this led to several problems. In particular, many old companies were recorded as new ones and this introduced anomalous counts as displayed in Fig. 2 left panel.

We use the Poisson generalized linear model to fit the number of new individual companies (SPISC) with respect to the total registered individual companies (SPREG), the trimester (TRIM), and the four provinces (PROV) of the Friuli Venezia Giulia region: Gorizia (GO), Pordenone (PN), Trieste (TS), and Udine (UD). Table 2 reports the estimates (Est.), the estimated standard errors (S.E.), and the $p$-values (P-val.) for three different methods: maximum likelihood (MLE) using function glm in the R package stats; the Ronchetti and Cantoni’s robust procedure (ROB) using function glmrob in the R package robustbase; and the weighted likelihood (WLE) using function wle.glm available in the R package wle. The SPREG is statistically insignificant for both MLE and ROB, while it is statistically significant for WLE. The second trimester (TRIM2) is not significant for both ROB and WLE, while the Gorizia province (PROVGO) is not significant for WLE, while it is highly significant for both MLE and ROB. We also notice that the sign of the fourth trimester coefficient (TRIM4) is the opposite in the robust procedures with respect to the MLE.
Table 1
Movimprese dataset. Estimated coefficients (Est.), standard errors (S.E.), and \(p\)-values (P-val.) for the model \(\text{SPISC} \sim \text{PROV-1}\) and three different methods. The column of MLE good shows the results of applying MLE on a dataset where observations with weights smaller than 0.1 are removed according to WLE-A.

|       | MLE |       |       | WLE |       | WLE_A |       |       |
|-------|-----|-------|-------|-----|-------|-------|-------|-------|
|       | Est. | S.E.  | P-val.| Est. | S.E.  | P-val.| Est.  | S.E.  | P-val. |
| PROVGO| 0.8473 | 0.0845 | 0.0000 | 0.2728 | 0.1167 | 0.0194 | 0.3215 | 0.1132 | 0.0045 |
| PROVPN| 2.4189 | 0.0385 | 0.0000 | 1.3541 | 0.0691 | 0.0000 | 1.4202 | 0.0657 | 0.0000 |
| PROVTS| −0.8755 | 0.2000 | 0.0000 | −0.9531 | 0.2088 | 0.0000 | −0.9271 | 0.2060 | 0.0000 |
| PROVUD| 2.6198 | 0.0348 | 0.0000 | 1.6297 | 0.0646 | 0.0000 | 1.7516 | 0.0586 | 0.0000 |
|       | 0.33898 | 0.11180 | 0.00243 | 1.45821 | 0.06389 | 0.00000 | −0.87547 | 0.19997 | 0.00001 |
|       | 1.92756 | 0.05143 | 0.00000 |
Table 2
Estimated coefficients (Est.), standard errors (S.E.), and $p$-values (P-val.) for the model $SPISC \sim PROV + TRIM + SPREG - 1$ and three different methods. The column of MLE good displays the results of applying MLE on a dataset where observations with weights smaller than 0.1 are removed according to WLE. Boldface highlights nonsignificant variables.

|          | MLE       | ROB       | WLE       | MLE good  |
|----------|-----------|-----------|-----------|-----------|
|          | Est. | S.E.  | P-val. | Est. | S.E.  | P-val. | Est. | S.E.  | P-val. | Est. | S.E.  | P-val. |
| PROVGO   | 1.004 | 0.094 | 0.0000 | 0.554 | 0.130 | 0.0000 | 0.165 | 0.173 | 0.339 | 0.362 | 0.140 | 0.00969 |
| PROVPN   | 2.600 | 0.091 | 0.0000 | 1.563 | 0.145 | 0.0000 | 0.925 | 0.243 | 0.410 | 0.194 | 0.253 | 0.0006 |
| PROVTS   | −0.729 | 0.204 | 0.0003 | −0.711 | 0.234 | 0.0024 | −0.974 | 0.265 | 0.0002 | −1.033 | 0.257 | 0.0006 |
| PROVUD   | 2.822 | 0.125 | 0.0000 | 1.884 | 0.200 | 0.0000 | 0.779 | 0.385 | 0.043 | 0.950 | 0.299 | 0.0016 |
| TRIM2    | −0.499 | 0.074 | 0.0000 | −0.117 | 0.097 | 0.215 | −0.197 | 0.139 | 0.157 | −0.042 | 0.107 | 0.69785 |
| TRIM3    | −0.923 | 0.086 | 0.0000 | −0.497 | 0.108 | 0.0000 | −0.527 | 0.149 | 0.0004 | −0.476 | 0.121 | 0.00009 |
| TRIM4    | 0.376 | 0.060 | 0.0000 | −0.813 | 0.120 | 0.0000 | −0.908 | 0.165 | 0.0000 | −0.790 | 0.134 | 0.0000 |
| SPREG    | −0.001 | 0.002 | 0.6210 | 0.0003 | 0.2150 | 0.0005 | 0.0040 | 0.0013 | 0.0039 | 0.0086 |
Table 3 displays the results of three different model selection procedures: Akaike Information Criterion (AIC), Weighted Akaike Information Criterion (WAIC), and Weighted Akaike Information Criterion based on Anscombe residuals (WAIC<sub>A</sub>). The column of AIC good reports the AIC for a dataset where observations with weights smaller than 0.1 are removed according to WLE at the full model.

Table 3

| Models            | AIC     | WAIC    | WAIC<sub>A</sub> | AIC good |
|-------------------|---------|---------|------------------|----------|
| SPISC~PROV+TRIM+SPREG | 4499.6  | 506.3   | 762.9            | 741.8    |
| SPISC~TRIM+SPREG   | 4960.0  | 547.9   | 827.1            | 817.0    |
| SPISC~PROV+SPREG   | 4856.9  | 537.3   | 814.5            | 787.8    |
| SPISC~SPREG        | 5306.7  | 581.9   | 881.9            | 865.8    |
| SPISC~PROV+TRIM    | 4497.8  | 514.0   | 775.1            | 752.4    |
| SPISC~TRIM         | 5825.2  | 772.5   | 1141.1           | 1113.1   |
| SPISC~PROV         | 4854.9  | 541.8   | 822.6            | 794.9    |
| SPISC~1            | 6182.3  | 793.4   | 1187.2           | 1151.3   |

Table 3 displays the results of three different model selection procedures: Akaike Information Criterion (AIC), Weighted Akaike Information Criterion (WAIC), and Weighted Akaike Information Criterion with weights based on Anscombe residuals (WAIC<sub>A</sub>) applied to the eight sub-models. The table is ordered in the index of WAIC that provides the same ranks as the WAIC<sub>A</sub>. The best model according to the WAIC is the full model, while for AIC the SPREG variable is dropped.

For further investigation an analysis of deviance is performed and the results shown in Table 4. Both MLE (based on Pearson’s statistic) and ROB (based on quasi-deviance decomposition) give large p-values for the SPREG variable, however weighted Pearson’s statistic based on WLE is highly significant. All methods agree on the importance of the TRIM variable.

5. Monte Carlo Simulation

In this section, we present some partial results of extensive Monte Carlo simulation experiments. The complete results are available in the supplemental material. The experiments involve three models: Poisson, gamma, and inverse Gaussian. The main features are the sample size (70, 140, 280, 560), and the level of contamination (0%, 5%, 10%). Only vertical outliers are investigated, while leverage points are not considered here. In our experiments, we compare the performance of the three methods: maximum likelihood estimate (MLE), the Ronchetti and Cantoni’s robust estimate (ROB, when available), and the proposed weighted likelihood estimate (WLE). For all combinations of models and features, we run 5,000 replications using the R software (R Development Core Team, 2011).

In our simulation for the Poisson family \(Y \sim \text{Poisson}(\lambda | x'^T \beta)\) with log link, we define four independent explanatory variables \(X_i, i = 1, \ldots, 4\) uniformly distributed on the discrete set \(-3, -2, -1, 0, 1, 2, 3\). The vector of parameters \(\beta\) is set to \{-0.25, -0.25, 0, 0\} with the intercept term \(\beta_0 = 2\) where the dispersion parameter \(\phi\) is known and set at 1. The Anscombe residual (Cox and Snell, 1968) for the Poisson model is defined as \(\frac{1}{2}(y^{2/3} - (\mu - 1/6)^{2/3})/\mu^{1/6}\). A Poisson model with linear predictors (10, 10, 10, 10), intercept 0, and the contaminated mean value truncated at 200 is used to generate the
Table 4
ANOVA decomposition using maximum likelihood (MLE), Ronchetti and Cantoni’s robust approach (ROB), and weighted likelihood (WLE). The column of MLE good reports the ANOVA decomposition using MLE for a dataset where observations with weights smaller than 0.1 are removed according to WLE at the full model.

| Models                  | $\chi^2$ (MLE) Test | P-val. | Q.D. (ROB) Test | P-val. | Weighted $\chi^2$ (WLE) Test | P-val. | $\chi^2$ (MLE good) Test | P-val. |
|-------------------------|---------------------|--------|-----------------|--------|-----------------------------|--------|--------------------------|--------|
| SPISC ~ PROV + TRIM + SPREG - 1 | 0.200               | 0.6222 | 0.810           | 0.3680 | 9.640                       | 0.0019 | 12.640                   | 0.0004 |
| SPISC ~ PROV + TRIM - 1  | 363.100             | 0.0000 | 41.584          | 0.0000 | 33.861                      | 0.0000 | 48.474                   | 0.0000 |
outliers. For MLE and ROB, we, respectively, employ, the function `glm (stats package)` and `glmrob (robustbase package)` using default values. WLE is implemented in the function `wle.glm (wle package)` by using the Generalized Kullback–Leibler

Figure 4. Poisson model. Bias of estimators. Level of contamination 10%, sample size $n = 140$. 
Table 5  
Poisson model. Actual significance level for different Wald-type test for $\beta_3$ parameter. Nominal level: $\alpha = 0.05$

| Size | Lev. | MLE  | WLE-0.05 | WLE-0.05-Asy | ROB |
|------|------|------|----------|--------------|-----|
| 70   | 0%   | 0.051| 0.042    | 0.050        | 0.050|
|      | 5%   | 0.796| 0.039    | 0.052        | 0.086|
|      | 10%  | 0.880| 0.056    | 0.073        | 0.154|
| 140  | 0%   | 0.052| 0.040    | 0.052        | 0.053|
|      | 5%   | 0.856| 0.040    | 0.051        | 0.085|
|      | 10%  | 0.920| 0.048    | 0.055        | 0.176|
| 280  | 0%   | 0.051| 0.038    | 0.049        | 0.048|
|      | 5%   | 0.901| 0.047    | 0.060        | 0.112|
|      | 10%  | 0.959| 0.045    | 0.061        | 0.235|
| 560  | 0%   | 0.052| 0.044    | 0.051        | 0.053|
|      | 5%   | 0.948| 0.048    | 0.065        | 0.137|
|      | 10%  | 0.992| 0.054    | 0.072        | 0.364|

(Okl) residual adjustment function with parameter $\tau$. Three different settings are considered: (i) $\tau = 0.05$ (WLE-0.05), (ii) $\tau = 0.1$ (WLE-0.1), and (iii) weights $w_i(\beta)$ are obtained from the Anscombe residuals with smoothing parameter $h = 0.031$ and $\tau = 0.05$ (WLE-0.05 Asy).

We first perform our simulations for discrete explanatory variables. Here follows a description in terms of bias for the investigated methods. In case of 0% contamination, MLE, ROB, and WLE-0.05-Asy are unbiased for all sample sizes. WLE-0.1 and WLE-0.05 are unbiased for $\beta_3$ and $\beta_4$, but exhibit slight bias for $\beta_1$ and $\beta_2$, and substantial bias for the intercept. In case of contamination, MLE breaks down, however, ROB shows bias for $\beta_1, \ldots, \beta_4$ with an almost unbiased intercept. WLE-0.1 and WLE-0.05 display stability; however, the biases for $\beta_1$ and $\beta_2$ are slightly increasing while the intercept is very biased. WLE-0.05-Asy performs well for all sample sizes and levels of contamination, with no bias for $\beta_1, \ldots, \beta_4$ and a slight bias on the intercept. Figure 4 shows the boxplots of the bias distributions for 10% contamination level and 140 sample size, which summarizes well the performance of the methods.

In Table 5, we report the actual significance level of a Wald-type test for the null hypothesis $\beta_3 = 0$ against the alternative hypothesis $\beta_3 \neq 0$ when the nominal significance level is fixed at 5%. The MLE is useless in the presence of contamination for all sample sizes; the ROB methods behave relatively well for low contamination levels and become quite distorted for moderate contamination levels, in particular that distortion increases with the sample size; the WLE in both versions shows high stability of the level for all sample sizes. A very similar pattern is observed for the other nominal levels, we have investigated: 0.01, 0.025, 0.1. The WLE-0.05 is slightly conservative while the WLE-0.05-Asy is slightly liberal.

In Fig. 5, we report the estimated distribution of the $p$-values on testing significance between the null hypothesis $H_0 : \beta_3 = 0$ and $\beta_4 = 0$ against the alternative hypothesis $H_1 : \beta_3 \neq 0$ and/or $\beta_4 \neq 0$ using different deviance tests, for sample size 140. Since we sampled from the null hypothesis, we expect a uniform distribution of the $p$-values. The density of the $p$-values are estimated using a Gaussian copula-based density estimator.
introduced in Jones and Henderson (2007) with bandwidth equal to 0.3 for all the cases. While the performance of the methods is good and equivalent in absence of contamination, the behavior is very different in the presence of contamination: the density of the \( p \)-values corresponding to MLE is concentrated in a right neighborhood of zero (not displayed in the plot) while the density of the \( p \)-values for the robust quasi-deviance test (ROB method) is highly nonuniform. While the WLE-0.05 and WLE-0.05-Asy have almost uniform \( p \)-value distribution with better performance of WLE-0.05-Asy when no contamination present. Similar behavior is observed for all sample sizes.

In our simulation for the gamma family \( Y \sim \text{gamma}(\mu(x^T \beta), \phi) \) with log link, we defined four independent explanatory variables \( X_i, i = 1, \ldots, 4 \), which are uniformly distributed on the discrete set \{1, 2, 3\}. The vector of parameters \( \beta \) is set to \{1, 1, 0, 0\} with the intercept term \( \beta_0 = 1 \) and the dispersion parameter \( \phi = 1 \). For this experiment, in addition, we consider sample sizes 1, 120 and 2, 240, and 15% of contamination. The Anscombe residual for the gamma model (McCullagh and Nelder, 1989) is defined as \( 3 \left( y^{1/3} - \mu^{1/3} \right) / \mu^{1/3} \). Outliers are generated by the following process, when \( \mu < 200 \) then the response variable is sampled from \( |N(3, 500, 25)| \) while for \( \mu \geq 200 \) the response variable is sampled from \( |N(0, 1, 0.0001)| \). For MLE and ROB, we, respectively employ, the function \texttt{glm} (stats package) and \texttt{glmrob} (robustbase package) using default values. WLE is implemented in the function \texttt{wle.glm} (wle package) by using the Generalized Kullback–Leibler (GKL) residual adjustment function with parameter \( \tau = 0.1 \). The following settings are used. The estimation of the dispersion parameter would be either based on weighted method of moments estimates (WLE-0.1-moment, see Eq. 1) or on weighted likelihood estimator (WLE-0.1-likelihood); weights based on the model and weights based on the Anscombe residuals. The folded standard normal distribution is used as kernel in the nonparametric kernel density estimators and in the evaluation of the smoothed model. The smoothing parameter is set iteratively to \( k \hat{\mu} \hat{\phi} \), where the constant \( k \) is set heuristically to: 1.024, 0.512, 0.256, 0.128, 0.064, 0.032 (smooth) according to the sample size for the usual weights. We further consider constants equal to \( k/2 \) (smooth/2) again according to sample size. For the asymptotic weights, a standard normal is used as kernel with bandwidth iteratively fixed to \( k\hat{\sigma} \); \( k \) set to 0.031 (asmooth/2) and to 0.062 (asmooth) regardless of the sample size and \( \hat{\sigma} \) is the actual estimates of the standard deviation; while the smoothed model is a normal model with zero mean and standard deviation equal to \( (1 + b)\hat{\sigma} \).
For all sample sizes and no contamination most of the considered methods are equivalent. The WLE based on likelihood and asymptotic Anscombe residuals (WLE 0.1 asymptotics, asmooth/2, likelihood) and the Cantoni and Ronchetti method (ROB) perform equivalently in most cases and they have the best performance. The WLE based on likelihood and asymptotic Anscombe residuals with double bandwidth (WLE 0.1 asymptotics, asmooth, likelihood) show a deterioration on the estimation of $\phi$. The WLE based on method of moments (WLE 0.1 smooth moment and WLE 0.1 smooth/2 moment) for the estimation of $\phi$ always shows a (stable over contamination level) bias for this parameter, besides that, the performance is very good. The WLE based on likelihood and bandwidth equal to smooth/2, most of the time has the worst performance, with large bias corresponding to parameters $\beta_1$, $\beta_2$, and $\phi$. In the use of Wald test, all methods have a good performance, however WLE based on the method of moments and Anscombe residuals (WLE 0.1 smooth, moment and WLE 0.1 smooth/2, moment) show a reduced power as the sample size grows, which make them useless for large sample sizes and all levels of contamination. In the ANOVA decomposition, based on Chi-square distribution, the WLE based on the likelihood and Anscombe residuals (WLE 0.1 asmooth/2, likelihood) performs best for all sample sizes and contamination levels. The Cantoni and Ronchetti method (ROB) performs equally well only for small sample sizes since, starting from sample size 280 in the presence of contamination its performance deteriorated significantly. WLE based on likelihood (WLE 0.1 smooth, likelihood and WLE 0.1 smooth/2, likelihood) shows a quite stable performance, however their distribution is not uniform under contamination, a similar behavior is observed for WLE 0.1 asmooth, likelihood. All types based on the method of moments do not perform very well. Very similar results are obtained using an $F$-type test in the ANOVA. In this case, ROB start to deteriorate slightly later and in relation with the contamination level: 560, 15%; 1, 120, 10%; and 2, 240, 5%.

In the third experiment, we consider the inverse Gaussian family, $Y \sim IG(\mu(x' \beta), \phi)$ with link: $1/\mu^2$. The predictors are four variables distributed as a discrete uniform on $(1, 2, 3)$. The vector of parameters $\beta$ is set equal to $\beta = (0.6, 0.6, 0, 0)$ with the intercept $\beta_0 = 0.01$ and the dispersion parameter $\phi = 0.01$. In this experiment, we consider sample sizes 70, 140, 280, 560 and the outliers are generated from an inverse Gaussian distribution with $\beta = (0, 0, 0, 0)$ and dispersion parameter equal to 10 with level of contamination set to 0%, 5%, 10%. The Anscombe residual for the gamma model (McCullagh and Nelder, 1989) is defined as $(\log(y) - \log(\mu))/\sqrt{\mu}$. For MLE, the function `glm` (stats package) is used while WLE is implemented in the function `wle.glm` (wle package) by using the Generalized Kullback–Leibler (GKL) residual adjustment function. Settings similar to those used for the gamma family are considered here.

The best results for estimation and inference seem to be achieved by weighted likelihood approach using the weighted method of moments for the estimation of dispersion parameter, regardless of the bandwidth (WLE 0.1, smooth/2, moment and WLE 0.1, smooth, moment). The performance of the WLE with dispersion parameter based on weighted likelihood is not very good, especially for small sample sizes. In particular, WLE with bandwidth “smooth” is the best, while the use of a smaller bandwidth (smooth/2) leads to worse results.

In case of no contamination and for all sample sizes, all the studied versions of WLE perform very well apart from WLE 0.1, smooth/2, likelihood where a bias is observed in the estimation of $\beta_1$ and $\beta_2$. Under contamination WLE 0.1, smooth/2, moment and WLE 0.1, smooth, moment perform best, while WLE 0.1, asmooth/2, moment and WLE 0.1, asmooth, moment show a loss of efficiency in the estimation of the $\beta$’s parameters and a moderate bias in the estimation of the dispersion parameter.
To study the behavior of the Weighted Wald test, we evaluate the statistics on evaluating the $\beta_3$ parameter. The WLE 0.1, smooth/2, moment and WLE 0.1, smooth, moment have shown a level close to the nominal one. The performance increases with sample size. To investigate the power of the Weighted Wald test, we test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Again, WLE 0.1, smooth/2, moment and WLE 0.1, smooth, moment show very high power regardless of the level of contamination and sample sizes.

In the use of likelihood ratio test for the analysis of variance (ANOVA), using both Chisquared-test and $F$-test, the WLE 0.1, smooth/2, moment and WLE 0.1, smooth, moment have a stable behavior with $p$-value distribution very close to the uniform distribution. The corresponding methods based on the Anscombe residuals (WLE 0.1, asmooth/2, moment and WLE 0.1, asmooth, moment) are useful only with moderate contamination (around 5%), while for larger contamination these last methods become unstable.

6. Conclusions

We have provided a general framework of robust estimation, inference, and model selection in a generalized linear model (GLM) for the exponential family. We have used and adapted the weighted likelihood method (WLE) to GLM. This method is highly related to the minimum distance/disparity methods and has the following main features: (i) outliers are defined on a “probabilistic” scale, rather than on a “geometrical” scale avoiding the location-scale formulation; (ii) WLE is a first-order asymptotic efficient estimates for the model of no contamination; and (iii) WLE has demonstrated good robust properties.

We have extended the method of Markatou et al. (1997, 1998) to generalized linear models by providing a complete robust inferential tool. We have introduced a modified version of Pearson residual function based on Anscombe residual, which is applicable to all types of datasets (with or without replicated observations) and the exponential family models. The Monte Carlo simulation results show the superior performance of our method compared to other robust approaches. Our method has been implemented by creating the function wle.glm, which is made available on CRAN in the R package wle. This function estimates and provides inferences for the binomial, Poisson, gamma, and inverse Gaussian models.

Appendix: Notes on Asymptotic and Bandwith Selection

A.1. Asymptotic Behavior of the Pearson Residuals Based on Anscombe Residuals

**Theorem A.1.** Let assume that (i) the kernel $k(z, t, h)$ is a bounded variation density kernel and the smoothing parameter $h$ is a positive constant; (ii) the observations $(y_i, x_i^j)$ $(i = 1, \ldots, n)$ come from a GLM with distribution $M(y, x_i^j; \theta_0, \phi_0)$, density $m(y, x_i^j; \theta_0, \phi_0)$ and $(\theta_0, \phi_0)$ belongs to the parameter space $(\Theta, \Phi)$. Let $r_i$ $(i = 1, \ldots, n)$ be the Anscombe residuals evaluated at the observation $(y_i, x_i^j)$ and parameter $(\theta_0, \phi_0)$ and let $G(r; \theta_0, \phi_0)$ be the distribution of the Anscombe residuals. The Pearson residuals based on the Anscombe residual $r$ are such that,

$$\sup_r |\delta_A(r; \theta_0, \phi_0)| \overset{a.s.}{\rightarrow} 0.$$
Proof. Let \( \tilde{f}(r) = \int k(r, t, h) d\tilde{F}_n(t) \) be the nonparametric kernel density estimate based on the \( r_i \) Anscombe residuals and \( \tilde{m}(r) = \int k(r; t, h) d\tilde{F}_v(t) \), where \( \tilde{F}_v(t) \) is an empirical distribution based on a sample of size \( v \) from \( G(r; \theta_0, \phi_0) \). We need to show that \( W_n = \sup_r |\tilde{f}(r) - \tilde{m}(z; \theta_0, \phi_0)| \xrightarrow{a.s.} 0 \) as \( n \to \infty \) and \( v \to \infty \). Let

\[
W_n = \sup_r \left| \int k(r, t, h) d\tilde{F}_n(t) - \int k(r, t, h) d\tilde{F}_v(t) \right|
\]

\[
= \sup_r \left| \int k(r, t, h) [\tilde{F}_n(t) - \tilde{F}_v(t)] \right|
\]

\[
= \sup_r \left\{ [k(r, t, h)]^{+\infty} - [k(r, t, h)]^{-\infty} \right\} d(\tilde{F}_n(t) - \tilde{F}_v(t))
\]

\[
< \xi \sup_t \left\{ [\tilde{F}_n(t) - G(t; \theta_0, \phi_0)] + [G(t; \theta_0, \phi_0) - \tilde{F}_v(t)] \right\},
\]

where \( \xi \) is the total variation of the kernel \( k \). The result follows from Glivenko-Cantelli Lemma. Then,

\[
\sup_r |\delta_A(r; \theta_0, \phi_0)| = \sup_r \left| \frac{\tilde{f}(r) - \tilde{m}(r; \theta_0, \phi_0)}{\tilde{m}(r; \theta_0, \phi_0)} \right| \xrightarrow{a.s.} 0 \quad n \to \infty, \quad v \to \infty.
\]

\[ \Box \]

A.2. Bandwidth Selection

When evaluating the weights for continuous models a smoothing parameter must be chosen. In finite sample, this parameter has an impact on the robustness performance of the procedure. Markatou et al. (1998) suggest the following procedure to determine the bandwidth when the underline model is normal and the kernel is normal as well. Consider the asymptotic form of the Pearson residual for a point mass contamination at \( c\sigma \) with contamination level \( \varepsilon \)

\[ \delta_{\varepsilon}(c) = \varepsilon \left( \left( \frac{k+1}{k} \right)^{1/2} \exp \left( \frac{c^2}{2(1+k)} \right) - 1 \right). \]

Inserting \( \delta_{\varepsilon}(c) \) in the weight function and equating it to a prefixed value \( w_{\text{a}} \) leads to \( k \). Consequently, the bandwidth is set to \( h = k \hat{\sigma} \), where \( \hat{\sigma} \) is an estimate of the scale parameter, so that the bandwidth is updated in each iteration. Using this approach we have \( k = 0.0031 \) for weight \( w_{\text{a}} \) equal to 0.03, level of contamination \( \varepsilon = 0.1 \) and position of the outlier at \( c = 4 \) for the GKL RAF and \( \tau = 0.1 \). Notice that the value of \( k \) is independent of the parameters \( \mu \) and \( \sigma \) of the normal distribution. The above approach is used when Anscombe residuals are considered in all our examples and simulations.

For other continuous models, such as gamma or inverse Gaussian, similar results could be obtained, however, the results will depend on the unknown parameters. For a given distribution \( F \) taking values on the positive line, location \( \mu \) and scale \( \sigma \), and kernel
$k(y, t, k\sigma)$, the asymptotic Pearson residual becomes

$$
\delta_a(c) = \varepsilon \left( \frac{k(c\sigma + \mu, c\sigma + \mu, k\sigma)}{\int k(c\sigma + \mu, t, k\sigma) \, dF_{\mu,\sigma}(t)} - 1 \right),
$$

which could be evaluated numerically for given values of $c$, $\varepsilon$, $\mu$, and $\sigma$, proceeding as above, we obtain a value for $k$. For instance, for the gamma distribution with shape $\alpha = 3$, scale $\beta = 1$, $\mu = \alpha\beta$, $\sigma = \sqrt{\alpha\beta}$, $c = 6$, $\varepsilon = 0.1$, and $w_a = 0.1$ lead to $k = 0.024$, while for the inverse Gaussian distribution with location $\mu = 1$, $\sigma = \mu^{3/2}$, $c = 5$, $\varepsilon = 0.1$, and $w_a = 0.1$ lead to $k = 0.039$.

**Acknowledgments**

The authors greatly appreciate the valuable comments of the referees, which resulted in significant improvement of the article. The authors also thank Prof. Hossein M. Soroukh from Kuwait University for his help in editing this manuscript.

**Supplemental Material**

Supplementary data for this article can be accessed on publisher’s website.

**References**

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statistics & Probability Letters* 56:289–300.

Agostinelli, C. (2006). Notes on Pearson residuals and weighted likelihood estimating equations. *Statistics & Probability Letters* 76:1930–1934.

Agostinelli, C., Markatou, M. (2001). Test of hypotheses based on the weighted likelihood methodology. *Statistica Sinica* 11(2):499–514.

Bianco, A., Yohai, V. J. (1996). Robust estimation in the logistic regression model, in robust statistics. In: Rieder, H., ed. *Data Analysis and Computer Intensive Methods, Proceedings of the workshop in honor of Peter J. Huber of Lecture Notes in Statistics*. Vol. 109. New York: Springer-Verlag, pp. 17–34

Brant, R. (1987). Residual components in generalized linear models. *The Canadian Journal of Statistics* 15(2):115–126.

Cantoni, E., Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association* 96(455):1022–1030.

Carroll, R. J., Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B* 55(3):693–706.

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society, Series B* 50(2):225–265.

Cox, D. R., Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B* 30(2):248–275. Available at: http://www.jstor.org/stable/2984505.

Cressie, N., Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B* 46:440–464.

Cressie, N., Read, T. R. C. (1988). Cressie–Read Statistic. In: Kotz, S., Johnson, N. L., eds. *Encyclopedia of Statistical Sciences*. Supplementary Volume. New York: Wiley, pp. 37–39.

Croux, C., Flandre, C., Haesbroeck, G. (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters* 60(4):377–386.

Croux, C., Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis* 44:273–295.
He, X., Simpson, D. G., Portnoy, S. L. (1990). Breakdown of robustness of tests. *Journal of the American Statistical Association* 85:446–452.

Hosseinian, S. (2009). *Robust Inference for Generalized Linear Models: Binary and Poisson Regression*. Ph.D. dissertation, Switzerland: École Polytechnique Fédérale de Lausanne.

Jones, M. C., Henderson, D. A. (2007). Kernel-Type density estimation on the unit interval. *Biometrika* 94(4):977–984.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association* 84:460–466.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics* 22:1018–1114.

Loynes, R. M. (1969). On Cox and Snell’s general definition of residuals. *Journal of the Royal Statistical Society, Series B* 31(1):103–106. Available at: http://www.jstor.org/stable/2984331.

Markatou, M., Basu, A., Lindsay, B. G. (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference* 57:215–232.

Markatou, M., Basu, A., Lindsay, B. G. (1998). Weighted likelihood estimating equations with a bootstrap root search. *Journal of the American Statistical Association* 93:740–750.

McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman and Hall.

Muller, C. H., Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* 116(2):503–519. Available at: http://www.sciencedirect.com/science/article/B6V0M-4724PXS-2/1/186842718ba15645e610b26ed5d682cf.

Park, C., Basu, A. (2003). The generalized Kullback–Leibler divergence and robust inference. *Journal of Statistical Computation and Simulation* 73(5):311–332.

Pierce, D. A., Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association* 81(396):977–986.

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

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org/.

Rousseeuw, P. J., Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis* 43(3):315–332. Available at: http://www.sciencedirect.com/science/article/B6V8V-4724PXS-2/1/8312cbe5989b01b3c8b95c9f92bd577f.

Ruckstuhl, A. F., Welsh, A. H. (2001). Robust fitting of the binomial model. *The Annals of Statistics* 29(4):1117–1136.

Stefanski, L. A., Carroll, R. J., Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* 73:413–424.