The Annals of Statistics
2001, Vol. 29, No. 4, 1117–1136

ROBUST FITTING OF THE BINOMIAL MODEL

BY A. F. RUCKSTUHL AND A. H. WELSH

Zürich University of Applied Sciences and University of Southampton

We consider the problem of robust inference for the binomial(m, π) model. The discreteness of the data and the fact that the parameter and sample spaces are bounded mean that standard robustness theory gives surprising results. For example, the maximum likelihood estimator (MLE) is quite robust, it cannot be improved on for $m = 1$ but can be for $m > 1$. We discuss four other classes of estimators: M-estimators , minimum disparity estimators, optimal MGP estimators, and a new class of estimators which we call E-estimators. We show that E-estimators have a non-standard asymptotic theory which challenges the accepted relationships between robustness concepts and thereby provides new perspectives on these concepts.

1. Introduction. Suppose that we have n observations $Y_1, \ldots, Y_n$ which are independent, take on values in the set $\{0, 1, \ldots, m\}$ and have a distribution which is plausibly close to the binomial(m, π) distribution with probability function

$$p_\pi(k) = \binom{m}{k} \pi^k (1 - \pi)^{m-k}, \qquad k = 0, \ldots, m,$$

where m is known and $0 < \pi < 1$ is an unknown parameter. In this paper, we study different estimators of π and their robustness properties under departures from the binomial model.

The problem of robust estimation of π in the binomial model has received little attention. Robust estimation in logistic regression has been considered by Pregibon [14], Stefanski, Carroll and Ruppert [18], Copas [4], Künsch, Stefanski and Carroll [10], Carroll and Pederson [2], Bianco and Yohai [1], Christmann [3], and Markatou, Basu and Lindsay [12] but has features which make it very different from the simple binomial problem. The literature on robust estimation for general models for independent identically distributed discrete data which is relevant to the binomial model treats M-estimators (Simpson et al. [17]), minimum disparity estimators (Simpson [16], Lindsay [11], and Markatou et al. [12]), and optimal MGP estimators of Victoria-Feser and Ronchetti [19] (i.e., an optimal generalized MLE with grouped data). The minimum disparity estimators are not first order robust; optimal MGP estimators are first order robust but have not been applied to the binomial model.

We propose to study robust estimation for the binomial model for independent identically distributed data in its own right rather than as an illustration of a general methodology. General methodologies for robustness are typically

Received April 1999; revised March 2001.

AMS 2000 *subject classifications*. Primary 62F12, 62F35.

Key words and phrases. Bias, breakdown point; E-estimation, influence function; likelihood disparity; M-estimation, minimum disparity estimation, optimal MGP estimation.

derived for problems with unbounded sample and parameter spaces so the intuition, implicit assumptions and results do not always successfully translate to the binomial model. We will show that robust estimation for the binomial model is an interesting problem with features which throw light on robustness concepts. Our study of this simple case allows us to clarify the issues and establish a basic framework for approaching problems which allow unequal m and incorporate covariates.

We discuss maximum likelihood estimation (MLE) under the binomial model in Section 2. We introduce a "gross error" contamination model (which can give rise to overdispersion) and study the robustness properties of the binomial MLE when this contamination model holds. Let $f_n(k)$ denote the proportion of observations equal to k in the sample of size n so

$$f_n(k) = n^{-1} \sum_{i=1}^{n} I(Y_i = k), \qquad k = 0, \ldots, m,$$

with I the indicator function. In Section 3, we consider the class of E-estimators $\widehat{\pi}$ of π which minimize $H(\pi, f_n)$, where

$$H(\pi, f_n) = \sum_{k=0}^{m} \rho\left(\frac{f_n(k)}{p_\pi(k)}\right) p_\pi(k) \tag{1.1}$$

and

$$\rho(x) = \begin{cases} (\log(c_1) + 1)x - c_1, & \text{if } x < c_1, \\ x\log(x), & \text{if } c_1 \le x \le c_2, \\ (\log(c_2) + 1)x - c_2, & \text{if } x > c_2. \end{cases}$$

When $c_1 = 0$ and $c_2 \to \infty$, $\hat{\pi}$ is the minimum relative entropy estimator which is identical to the MLE for the binomial model. We are particularly interested in the choice $c_2 = 1$ which leads to attractive robust estimators. In Section 4, we explore the relationship of E-estimators to M-estimators, minimum disparity estimators, and optimal MGP estimators. The asymptotic properties of E-estimators are presented in Section 5. We compute the influence function and show that the choice $c_1 < c_2 = 1$ improves on the robustness of the available estimators under "gross error" contamination. Then we derive the asymptotic distribution of E-estimators.

2. Maximum likelihood estimation and contamination. Suppose first that we have a sample of n independent and identically distributed observations $Y_1, \ldots, Y_n$ which are generated by a binomial(m, π) model with true parameter $\pi = \pi_*$. The log-likelihood is $n^{-1}\sum_{i=1}^{n}\log\{p_\pi(y_i)\} = \sum_{k=0}^{m}\log\{p_\pi(k)\} \times f_n(k)$ so the estimating equation is

$$0 = n^{-1} \sum_{i=1}^{n} s(y_i, \widehat{\pi}_{\mathrm{MLE}}) = \sum_{k=0}^{m} s(k, \widehat{\pi}_{\mathrm{MLE}}) f_n(k), \tag{2.1}$$

where

$$s_\pi(k) = s(k, \pi) = (k - m\pi)/(\pi(1 - \pi))$$

is the score function for the binomial model. The MLE is $\widehat{\pi}_{\text{MLE}} = \bar{y} = 1/m \sum_{k=0}^{m} k f_n(k)$, and the Fisher information at the binomial(m, π) model is $I(\pi) = \sum_{k=0}^{m} s_\pi(k)^2 p_\pi(k) = m/\pi(1-\pi)$.

Next, suppose the observations $Y_1, \ldots, Y_n$ do not follow a binomial model exactly. The usual approach in robustness is to assume that the underlying distribution has a probability function $f(k)$ which is close to a binomial distribution in the sense that for some $0 < \pi_* < 1$,

$$f(k) = (1-\gamma) p_{\pi_*}(k) + \gamma\, q(k), \qquad k = 0, \ldots, m, \tag{2.2}$$

where q is an arbitrary probability function on $\{0, \ldots, m\}$ and $0 \le \gamma \le 1$. Contamination models like (2.2) can give rise to overdispersion: if q is conditionally binomial(m, u) given u, where u has a beta$(\pi_*(1/\tau^2 - 1),\ (1-\pi_*)(1/\tau^2 - 1))$ distribution, the mean and variance under f are

$$\begin{aligned} \mathrm{E}_f(Y) &= m\pi_*, \\ \mathrm{Var}_f(Y) &= \{1 - \gamma + \gamma(m-1)\tau^2\} m \pi_* (1-\pi_*), \end{aligned}$$

which is similar to the beta-binomial model for overdispersion (see McCullagh and Nelder [13], Section 4.5 and page 140). From a robustness point of view, overdispersion results from a special form of contamination in which there is no bias and only an effect on the variance. Model (2.2) allows for more general forms of contamination (inflation of arbitrary classes, in the tails or otherwise) so we cannot treat all contaminated binomial data by the simple expedient of incorporating an overdispersion parameter into the model. We will focus on the contamination model (2.2) but note that other types of "contamination" not covered by model (2.2) (such as censoring leading to an unobserved zero class) can occur.

An important point is that there can be no contamination (and hence no overdispersion) when $m = 1$. In this case q must be a binary distribution so the linear combination f of p_{π_*} and q is also binary. This also means that all reasonable estimators are identical to the MLE when $m = 1$.

One way to study the stability of an estimator under contamination is to determine the breakdown point ([6], [8]). The MLE breaks down when it is at the edge of the parameter space (i.e. equals zero or one) and this occurs only when all the observations equal 0 or m. It follows that the MLE has breakdown point bounded by $1 - \max(f_n(0), f_n(m))$. For large m, and $\pi_* \approx 1/2$, the bound tends to one. While at first impressive, this high breakdown property can be quite misleading. In particular, if the data are actually generated by an equal mixture of two binomial distributions, the MLE estimates the average $(\pi_1 + \pi_2)/2$ which can be far from a sensible answer even though the estimator has not broken down.

Under the contamination model (2.2), the MLE is estimating

$$\pi_o = \frac{1}{m} \sum_{k=0}^{m} k f(k) = (1-\gamma)\pi_* + \gamma \frac{\mu_q}{m}$$

so is a biased estimator of π_* whenever $\gamma > 0$ and $\mu_q \neq m\pi_*$. For general estimators the bias is nonlinear, but when $q(k)$ in the contamination model (2.2) is chosen to be the pointmass 1 at any x, we can approximate the bias by the influence function of the estimator. The influence function of the MLE for the binomial model at x is

$$IF(x, \widehat{\pi}_{\mathrm{MLE}}, p_{\pi_*}) = I(\pi_*)^{-1} s_{\pi_*}(x) = \frac{x}{m} - \pi_*, \qquad x = 0, 1, \ldots, m. \tag{2.3}$$

The gross error sensitivity is

$$\max_x \left| IF(x, \widehat{\pi}_{\mathrm{MLE}}, p_{\pi_*}) \right| = \max(\pi_*, 1 - \pi_*) < \infty$$

(see [8], Section 2.1). Thus the MLE has bounded influence and can be regarded as a robust estimator. However, simply having bounded influence (in this case, because the sample space is bounded) does not imply that the estimator is a good robust estimator or cannot be improved on. The robustness problem for the binomial model is one of trying to find a robust estimator which improves on the MLE in the sense, say, of having smaller bias under contamination or smaller gross error sensitivity.

3. *E*-estimators. A central idea in the development of methodology to handle gross-error contamination is to modify the log-likelihood to reduce the effect of observations in the tails of the distributions. However, we are concerned with the effects of distributional contamination which is not restricted to the tails and which may tend to deflate as well as inflate various classes in arbitrary ways. This suggests that it is more sensible to work on the frequency scale than on the scale of the individual observations. The most natural way to decide which classes to control is through relating the relative frequencies to the probability function of the assumed binomial model, an insight which leads us toward minimum distance estimation on the relative frequency scale.

The choice of divergence or disparity to work from is initially rather arbitrary though ultimately is confirmed by the properties and performance of the estimator. It makes sense to work from a disparity which produces the MLE when the binomial model holds. We will work with the *likelihood disparity* $H(\pi, f_n)$, where

$$H(\pi, f_n) = \sum_{k=0}^{m} \rho\left(\frac{f_n(k)}{p_\pi(k)}\right) p_\pi(k),$$

with $\rho(x) = x\log(x)$. (Note that the likelihood disparity is the relative entropy of f_n with respect to p_π and is one of the forms of Kullback-Leibler divergence.) The likelihood disparity is minimized by the MLE $\hat{\pi}_{\mathrm{MLE}}$ for the binomial model and hence bears the same relationship to the log-likelihood as least squares to maximum likelihood in Gaussian models. Just as Huber [9] replaced the quadratic function away from the turning point by a linear function, we replace $x\log(x)$ by a linear function away from one. This has the effect of reducing the rate at which $\rho(x)$ tends to infinity and reduces the effect of classes for which $f_n(k)/p_\pi(k)$ is large. We can also act on classes for which $f_n(k)/p_\pi(k)$ is small

by preventing $x\log(x)$ from tending to zero as x decreases to zero. That is, we also replace $x\log(x)$ by a linear function when x is smaller than one. Imposing the requirement that the modified ρ has a continuous derivative, we are led to the objective function (1.1).

Since $f_n \to f$ almost surely and $H(\pi, f)$ is continuous at f, the objective function $H(\pi, f_n)$ is estimating $H(\pi, f)$. The precise shape of $H(\pi, f)$ depends on both f and the values of c_1 and c_2. When $c_2 < \infty$, H is a bounded function of π but otherwise is unbounded at zero and/or one. The objective function H is also generally nonconvex in π and, as we will show below, for $c_1 > 0$ and/or $c_2 < \infty$ not smooth in the sense that the second derivative does not exist for all π.

When $c_1 = 0$ and $c_2 = \infty$, the MLE can be obtained explicitly so H always has a unique minimum. However H can have multiple local minima when $0 \le c_1 < 1$ and/or $c_2 < \infty$ and $\gamma \ne 0$ in (2.2). In particular, if $c_1 = 0$ and $c_2 < \infty$, $\gamma = 0.5$ and q is a different binomial distribution from p_{π_*}, then H has two minima, showing that the "breakdown point" of the estimator with $c_2 < \infty$ is at most 1/2. Multiple minima suggest that an underlying binomial model is too simple for the observed data. We will focus on cases in which the contamination q is such that $H(\pi, f)$ has a unique global minimum. That is,

$$\pi_o = \pi_o(f) = \operatorname*{argmin}_{\pi} H(\pi, f) \tag{3.1}$$

is a unique, interior point of $0 \le \pi \le 1$ and $H(\pi, f)$ is convex in a neighborhood of π_o. Since we have a bounded univariate parameter π, it is simple to plot $H(\pi, f)$ and we can obtain the minimum by direct searching. Clearly $\hat{\pi}$ is estimating π_o which identifies the binomial distribution which best approximates f when closeness is measured by H. The bias of the estimator $\hat{\pi}$ is given by $\pi_o(f) - \pi_* = \pi_o(f) - \pi_o(p_{\pi_*})$. The bias is zero when $f = p_{\pi_*}$ which occurs under model (2.2) when $\gamma = 0$. Generally, we have a nonzero bias so good procedures will be ones for which in the first instance π_o is close to π_* and in the second produce estimators with small variability.

With the weight function

$$w(x) = \rho'(x) - \rho(x)/x = I(c_1 \le x \le c_2) + \frac{c_1}{x} I(x < c_1) + \frac{c_2}{x} I(x > c_2), \tag{3.2}$$

the estimating equation for π is

$$0 = \eta(\widehat{\pi}, f_n), \tag{3.3}$$

where

$$\eta(\pi, f_n) = -\sum_{k=0}^{m} w\left(\frac{f_n(k)}{p_\pi(k)}\right) s_\pi(k) f_n(k).$$

Observed frequencies which are greater than $c_2 p_\pi(k)$ are downweighted while observed frequencies which are smaller than $c_1 p_\pi(k)$ are given additional weight. This is natural on the frequency scale and represents exactly what we intuitively require in our weight function. For $m = 1$, the E-estimator is identical to the MLE.

Note that the estimating function $\eta(\pi, f_n)$ is continuous in π but differentiable only at the values of π for which $\frac{f_n(k)}{p_\pi(k)} \neq c_1$ or c_2 for any $k = 0, 1 \ldots, m$. That is, the estimating function is differentiable except possibly at a finite set of points.

4. Relationship to other approaches.

4.1. *M-estimation.* The earliest approach to robustness for binomial data was presented by Hampel [7], page 95. He noted that the influence function of the MLE is bounded but argued that better estimators should be available for moderate or large m. In this case, working on the observation scale, he applied his Lemma 5 to obtain the M-estimator which satisfies

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \psi_b \left(\frac{Y_i - m\,\pi}{\sqrt{\pi(1-\pi)/m}} - a(\pi) \right) \\
&= \sum_{k=0}^{m} \psi_b \left(\frac{k - m\,\pi}{\sqrt{\pi(1-\pi)/m}} - a(\pi) \right) f_n(k),
\end{aligned} \tag{4.1}
$$

where $\psi_b(x) = x \min(1, b/|x|)$ is Huber's ψ function, b is constant for all π and the scalar $a(\pi)$ is chosen to satisfy the consistency condition

$$
0 = \sum_{k=0}^{m} \psi_b \left(\frac{k - m\,\pi}{\sqrt{\pi(1-\pi)/m}} - a(\pi) \right) p_\pi(k) \tag{4.2}
$$

so that the M-estimator is Fisher-consistent. It makes sense to use a smooth ψ-function such as the scaled logistic function in place of Huber's ψ-function in (4.1) (see [17]).

The influence function of the M estimator defined in (4.1) for any reasonable ψ-function is

$$
\mathrm{IF}(x, \widehat{\pi}_\psi, p_{\pi_*}) = c^{-1}(\pi_*) \psi \left(\frac{x - m\,\pi_*}{\sqrt{\pi_*(1-\pi_*)/m}} - a(\pi_*) \right)
$$

with

$$
c(\pi_*) = \sum_{k=0}^{m} \psi \left(\frac{k - m\,\pi_*}{\sqrt{\pi_*(1-\pi_*)/m}} - a(\pi_*) \right) s(k,\ \pi_*) p_{\pi_*}(k).
$$

This function is shown for Huber's ψ function in Figure 1. The gross error sensitivity is $c^{-1}(\pi_*) \sup_x |\psi(x)|$ which depends on π_*. Note that in contrast to the E-estimator (and the estimates in Section 4.2–4.3), the M-estimator is linear in f_n.

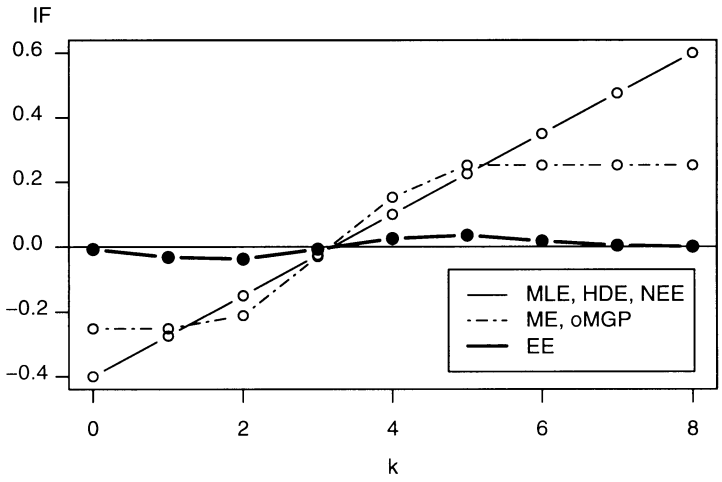

FIG. 1. *Influence functions for maximum likelihood estimator* (*MLE*), *Hellinger distance estimator* (*HDE*), *estimator with negative exponential RAF* (*NEE*), *M-estimator and optimal MGP-estimator using Huber's* ψ*-function with* $b = 1$ (*ME*) *and with* $b = 0.252$ (*oMGP*), *respectively, and E-estimator with* $c_1 = 0$ *and* $c_2 = 1$ (*EE*) *at* $\pi_* = 0.4$ *and* $m = 8$.

4.2. *Minimum disparity and weighted likelihood estimators.* A large class of estimators including the MLE, the minimum power divergence estimators ([5] and [15]), the minimum disparity estimators (Lindsay [11]) and the weighted likelihood equations estimators (Markatou et al. [12]) can be written as the solution of

$$0 = \sum_{k=1}^{m} w\left\{\frac{f_n(k)}{p_\pi(k)}\right\} s_\pi(k) f_n(k). \tag{4.3}$$

(Our notation is slightly different from that of Markatou et al. [12] because they use as the argument of w the quantity $(f_n(k)/p_\pi(k) - 1)$ instead of $f_n(k)/p_\pi(k)$.)

Markatou et al. [12] apply the results of Lindsay [11] to show that the minimum disparity/weighted likelihood equations estimators are asymptotically equivalent to the MLE when the binomial model holds. This means that these estimators have the same influence function and hence, in the sense of influence, the same robustness properties. Lindsay [11] and Markatou et al. [12] explore the next term in the von Mises expansion of the estimator. They show that large negative values of $w''(0)$ lead to second order robustness and that if $w(x) \sim x^{-0.5}$ as $x \to \infty$, the estimator has breakdown point 0.5 (see discussion in Section 2).

Equation (4.3) is of the same form as the estimating equation (3.3) for the E-estimator but with a different choice of weight function. Markatou et al. [12] concentrated on the Hellinger distance estimator which has $w(x) = (2x^{1/2} - 1)/x$ and the negative exponential estimator with $w(x) = (3 - (1 + x)\exp(1 - x))/x$, which should be compared to (3.2). The negative

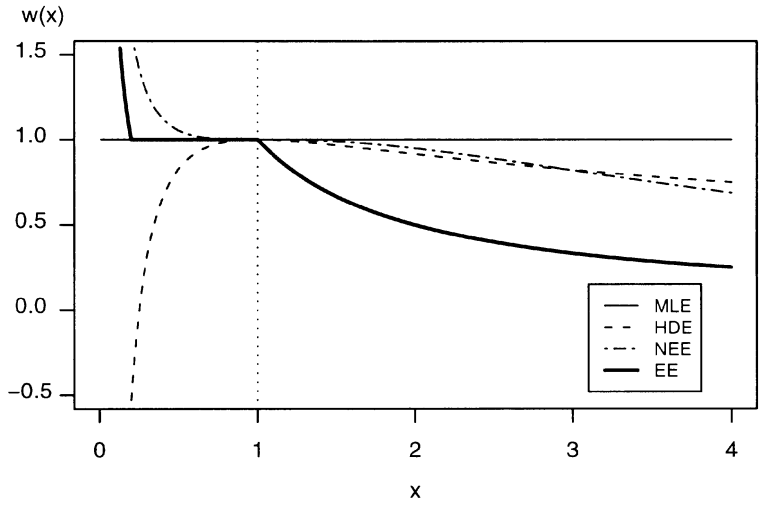

FIG. 2. *Weight functions $w(x)$ for maximum likelihood estimator (MLE), Hellinger distance estimator (HDE), estimator with negative exponential RAF (NEE), and E-estimator with $c_1 = 0.2$ and $c_2 = 1$ (EE). M- and MGP-estimators cannot be represented by a graph $w(x)$ versus x.*

exponential estimator produces a weight function which is similar in spirit to that of the E-estimator but, crucially, is much smoother and decreases more slowly for large x (see Figure 2). The weight function for the E-estimator is not smooth at $x = 1$ so the theory of Markatou et al. [12] does not apply.

4.3. *Generalized maximum likelihood estimator for grouped data.* Victoria-Feser and Ronchetti [19] considered robust estimation of the parameters of a continuous model from grouped data. The methodology can be viewed as a general approach to robustness for models for discrete data. For simplicity, we describe the methodology in the context of the binomial problem.

Starting from the class of minimum power divergence estimators, Victoria-Feser and Ronchetti [19] introduced the class of MGP estimators (generalized MLE for grouped data) which satisfy

$$0 = \sum_{k=0}^{m} \left\{ \frac{f_n(k)}{p_\pi(k)} \right\}^{\lambda+1} \{\psi_\pi(k) p_\pi(k)^\lambda\}\, p_\pi(k) = \sum_{k=0}^{m} \psi_\pi(k) f_n(k)^{\lambda+1}, \tag{4.4}$$

where $-\infty < \lambda < \infty$ is fixed. Clearly, if we let $w(k) = \psi_\pi(k) f_n(k)^\lambda / s_\pi(k)$, the estimator is of the same form as the minimum disparity/weighted likelihood equations estimators except that $w(k)$ is no longer a function of $f_n(k)/p_\pi(k)$ alone.

Victoria-Feser and Ronchetti [19] obtained the influence function of the MGP estimators and showed that the optimal MGP estimator (the estimator with smallest asymptotic variance subject to a bound on the gross error

sensitivity) has

$$\psi_\pi(k) = \left(\frac{1}{p_\pi(k)}\right)^\lambda \psi_b[G(\pi)\{s_\pi(k) - a(\pi)\}],$$

where $\psi_b(x) = x\min(1, b/|x|)$ is the Huber ψ-function, with $G(\pi)$ and $a(\pi)$ determined by the Fisher consistency condition

$$\sum_{k=0}^{m} \psi_b[G(\pi)\{s_\pi(k) - a(\pi)\}]p_\pi(k) = 0$$

and the normalization condition

$$\sum_{k=0}^{m} \psi_b[G(\pi)\{s_\pi(k) - a(\pi)\}]p_\pi(k)s_\pi(k) = 1.$$

The effect of the normalization condition is to make the influence function of the optimal MGP equal to the ψ_b-function and consequently the gross error sensitivity independent of the parameter π_*. This means also that the influence function of the optimal MGP estimator does not depend on λ and hence the optimal MGP estimators are asymptotically identical to the B-optimal estimator ([8], pages 243–244).

Victoria-Feser and Ronchetti [19] showed that the optimal MGP estimators have different small sample robustness properties with the optimally robust Hellinger distance estimator ($\lambda = -0.5$) performing better than the optimally robust MLE ($\lambda = 0$) in contaminated finite samples.

5. Some properties of the *E*-estimator. We derive the influence function and discuss the robustness properties of E-estimators in Subsection 5.1. We then study the asymptotic distribution of the E-estimator in Subsection 5.2 and show that, under the model, it is, in general, non-Gaussian.

5.1. *Influence function.* The asymptotic bias of $\widehat{\pi} = \pi_o(f_n)$ under the model f given by (2.2) is $\pi_o(f) - \pi_o(p_{\pi_*}) = \pi_o(f) - \pi_*$. The "change-of-bias" function can be obtained by dividing by γ and then letting $\gamma \to 0$. That is

$$CoB(p_{\pi_*}, q) = \lim_{\gamma\to 0} \frac{\pi_o(f) - \pi_*}{\gamma}.$$

If $c_1 < c_2 = 1$, then for $q(k) \neq p_{\pi_*}(k)$ and γ sufficiently small $f(k)/p_{\pi_o}(k) \neq c_1$ or c_2 so the weight function w is differentiable with respect to γ. If $c_2 > 1$, then it is possible that $f(k)/p_{\pi_o}(k) = c_1$ or c_2; in this case choose a smaller $\gamma > 0$ so that w is differentiable with respect to γ. Then differentiation of the estimating equation

$$0 = -\sum_{k=0}^{m} w\left(\frac{f(k)}{p_{\pi_o}(k)}\right) s_{\pi_o}(k)f(k)$$

with respect to γ yields

$$0 = -\sum_{k=0}^{m} w\left(\frac{f(k)}{p_{\pi_o}(k)}\right) s_{\pi_o}(k)\{q(k) - p_{\pi_*}(k)\} - \sum_{k=0}^{m} w\left(\frac{f(k)}{p_{\pi_o}(k)}\right) s'_{\pi_o}(k)\pi'_o f(k)$$
$$- \sum_{k=0}^{m} w'\left(\frac{f(k)}{p_{\pi_o}(k)}\right)\left\{\frac{q(k) - p_{\pi_*}(k)}{p_{\pi_o}(k)} - \frac{f(k)p'_{\pi_o}(k)\pi'_o}{p_{\pi_o}(k)^2}\right\} s_{\pi_o}(k) f(k).$$

If $c_1 < 1 < c_2$, the functions w and w' are continuous at 1 and satisfy $w(1) = 1$ and $w'(1) = 0$ respectively, so

$$CoB(p_{\pi_*}, q) = \sum_{k=0}^{m} IF(k, \widehat{\pi}_{\mathrm{MLE}}, p_{\pi_*})\, q(k),$$

where IF is given by (2.3), is linear in q and the influence function at x under the model (obtained by replacing q by the point mass at x) is also given by (2.3), confirming that $\hat{\pi}$ is fully efficient when $c_2 > 1$ and the binomial model holds. If $c_2 = 1$, $w'(1-) = 0$ and $w'(1+) = 1$ so

$$CoB(p_{\pi_*} q) = \frac{\sum_{k=0}^{m} s_{\pi_*}(k)\{q(k) - p_{\pi_*}(k)\} I[q(k) \le p_{\pi_*}(k)]}{\frac{m}{\pi_*(1-\pi_*)} - \sum_{k=0}^{m} s_{\pi_*}(k)^2 p_{\pi_*}(k) I[q(k) > p_{\pi_*}(k)]} \tag{5.1}$$

which is nonlinear in q. Taking the supremum over all q, Künsch pointed out that we obtain

$$\sup_q |CoB(p_{\pi_*} q)| = \frac{\pi_*(1-\pi_*)}{\min(\lceil m\pi_* \rceil - m\pi_*,\ m\pi_* - \lfloor m\pi_* \rfloor)},$$

where $\lfloor x \rfloor$ is the largest integer strictly smaller than x and $\lceil x \rceil$ is the smallest integer strictly larger than x. The supremum is achieved by putting $q(k) = 0$ for $k = \lfloor m\pi_* \rfloor$ or $k = \lceil m\pi_* \rceil$ respectively and $q(k) > p_*(k)$ for all other values of k. This is (partial) truncation contamination which is the antithesis of point contamination in the sense that an E-estimator with $c_1 = 1$ and $c_2 > 1$ rather than with $c_1 = 0$ and $c_2 = 1$ should be used for this kind of contamination (see discussion below).

If we consider point contamination at x and explore its effect, we obtain

$$\begin{aligned} IF(x, \widehat{\pi}, p_{\pi_*}) &= \frac{s_{\pi_*}(x) p_{\pi_*}(x)}{\frac{m}{\pi_*(1-\pi_*)} - s_{\pi_*}(x)^2 p_{\pi_*}(x)} \\ &= \frac{\pi_*(1-\pi_*)(x - m\pi_*) p_{\pi_*}(x)}{m\pi_*(1-\pi_*) - (x - m\pi_*)^2 p_{\pi_*}(x)}. \end{aligned} \tag{5.2}$$

We still refer to (5.2) as the influence function although it does not satisfy the relation $\sup_q |CoB(p_{\pi_*}, q)| = \sup_x |IF(x, \widehat{\pi}, p_{\pi_*})|$ which holds in standard robustness theory.

The influence functions of the MLE, M-estimator with $b = 1$, and the E-estimator with $c_1 = 0$ and $c_2 = 1$ at $\pi = 0.4$ and $m = 8$ are plotted in Figure 1 for comparison. If we plot the influence function at the binomial model as a function of x for different choices of m, π, γ, c_1 and c_2, we see that the choice

of c_1 has no effect on the influence function and that $c_2 = 1$ produces a more robust estimator than any other choice. But note that the point contamination is "the most favourable" contamination for an E-estimator with $c_1 < 1$ and $c_2 = 1$. In the worst case, which is identical to "the most favourable" contamination for an E-estimator with $c_1 = 1$ and $c_2 > 1$, the absolute "change-of-bias" is $\sup_q |CoB(p_{\pi_*}\, q)| = 1.2$.

The expression (5.2) is difficult to handle when computing the gross error sensitivity (which measures the greatest effect of pointmass contamination). An approximation to (5.2) can be achieved by replacing $p_{\pi_*}(x)$ by the standard normal density function ϕ at $(x - m\pi_*)/\sqrt{m\pi_*(1-\pi_*)}$. Maximising the numerator and minimising the denominator separately in the approximated influence function results in the approximate upper bound:

$$|IF_{approx}(x, \widehat{\pi}, p_{\pi_*})| \le \frac{\phi(1)}{1 - 2\phi(\sqrt{2})} \sqrt{\frac{\pi_*(1-\pi_*)}{m}} = 0.343 \sqrt{\frac{\pi_*(1-\pi_*)}{m}}.$$

For the example used in Figure 1 we obtain a value of 0.06 which is in close agreement with the maximum of the influence function of the E-estimator plotted there. This result can be compared to the lower bound of the gross error sensitivity of the optimal MGP estimators given in Proposition 4 of [19]. Calculating the lower bound for the binomial model we obtain

$$|IF(x, \widehat{\pi}_{oMGP}, p_{\pi_*})| \ge \left\{ \sum_{k=0}^{m} |s_{\pi_*}(k)| p_{\pi_*}(k) \right\}^{-1} \ge \sqrt{\frac{\pi_*(1-\pi_*)}{m}}$$

using Cauchy-Schwarz inequality. Hence the lower bound of the gross error sensitivity of the optimal MGP estimators is larger by a factor of almost 3 than the approximate upper bound of the absolute value of the influence function of E-estimators with $c_2 = 1$.

The lack of effect of c_1 in the influence function (5.2) is due to the fact that, under point contamination, the ratio $f(k)/p_\pi(k)$ is equal to $1-\gamma$ except when $k = x$ where we get $1 - \gamma + \gamma/p_\pi(k)$ which is larger than one. This means that the lower part of the weight function either has no effect or applies to nearly all the terms according to whether $1 - \gamma \ge c_1$ or $1 - \gamma < c_1$. Thus we might be inclined to take c_1 equal to or close to 0.

In other contamination models it makes sense to choose c_1 close to 1 so that cells with too small observed frequencies are "up-weighted" (cf. Figure 2). Suppose we observe a binomial(m, π_*) distribution truncated at zero. Then the relative frequency distribution will converge almost surely to

$$f_o(k) = \left(p_{\pi_*}(k) - \delta_0(k)\, p_{\pi_*}(k)\right) / (1 - p_{\pi_*}(0))$$

provided that $p_{\pi_*}(0) < 1$ and the E-estimator will estimate the parameter π_o which minimizes $H(\pi, f_o)$. If there is no local maximum between π_* and π_o, then we can study the slope of the tangent of $H(\pi, f_o)$ at $\pi = \pi_*$ to show on which side of π_* the parameter π_o lies. Since $f_o(0)/p_{\pi_*}(0) = 0$ and

$f_o(k)/p_{\pi_*}(k) > 1$ for any $k > 0$, for $0 \le c_1 < 1$ and $c_2 = 1$ we can write the slope as

$$H'(\pi_*, f_o) = \frac{\partial}{\partial \pi} H(\pi, f_o)\bigg|_{\pi=\pi_*} = (c_1 - 1)\, m\, (1 - \pi_*)^{m-1}.$$

Since $H'(\pi_*, f_o)$ is negative, the parameter π_o is larger than π_*. That is, using the E-estimator with $0 \le c_1 < 1$ and $c_2 = 1$ for the binomial(m, π_*) distribution truncated at zero, results in an estimate which is larger than it should be. If the linearisation is a good approximation to $H(\pi, f_o)$ for π between π_* and π_o, then a choice of c_1 close to 1 results in a smaller bias than a choice of c_1 close to zero and we can consider choosing $c_1 = 1$ and $c_2 > 1$. This is incompatible with choosing $c_1 < 1$ and $c_2 = 1$ and again shows the difficulty of trying to deal with two different types of contamination simultaneously. The analysis also shows that the MLE ($c_1 = 0$, $c_2 = \infty$) is less robust to truncation and Figure 2 shows that the Hellinger distance estimator is not robust to truncation (see also Section 7.2 in [11]).

5.2. *Asymptotic distribution.* We first establish that the E-estimator is consistent.

LEMMA 1. *Suppose that* $0 \le c_1 < 1$ *and* $c_2 \ge 1$, π_o *is the unique minimum of* $H(\pi, f)$ *in* $[0, 1]$ *and that* π_o *is an interior point of* $[0, 1]$. *Then* $\hat{\pi} \to \pi_o$ *almost surely*.

The proof of this lemma is given in the Appendix.

The optimal MGP estimators minimize (the trace of) their asymptotic variances subject to a bound on the gross error sensitivity which, however, must be larger than a lower limit (see [19]). We have argued in Subsection 5.1 that the E-estimator with $c_2 = 1$ has a lower gross error sensitivity than the lowest value possible for optimal MGP estimators. This indicates that the E-estimator has an unusual asymptotic distribution.

THEOREM 1. *Suppose that* $0 \le c_1 < 1$ *and* $c_2 \ge 1$, π_o *is the unique minimum of* $H(\pi, f)$ *in* $[0, 1]$, *that* π_o *is an interior point of* $[0, 1]$ *and that* H *is convex in a neighborhood of* π_o. *Let*

$$v(\mu) = \left(I\left(c_1 \le \frac{f(0)}{p_\mu(0)} \le c_2\right) s_\mu(0), \ldots, I\left(c_1 \le \frac{f(m)}{p_\mu(m)} \le c_2\right) s_\mu(m) \right)^T$$

and $\Sigma = diag(f) - ff^T$ *with* $f = (f(0), \ldots, f(m))^T$. *Then, for* $t \ne 0$,

$$\Pr\left(n^{1/2} \frac{\hat{\pi} - \pi_o}{V_t} \le t \right) \to \Phi(t)$$

with

$$V_t = \begin{cases} \lim_{\mu \downarrow \pi_o} V(\mu) = V(\pi_o+), & \text{if } t > 0, \\ \lim_{\mu \uparrow \pi_o} V(\mu) = V(\pi_o-), & \text{if } t < 0, \end{cases}$$

where

$$V^{-1}(\mu) = \sum_{k=0}^{m} \left\{ w' \left(\frac{f(k)}{p_\mu(k)} \right) \frac{f(k)}{p_{\pi_o}(k)} s_{\pi_o}(k)^2 - w \left(\frac{f(k)}{p_{\pi_o}(k)} \right) s'_{\pi_o}(k) \right\} f(k)/\sigma(\mu)$$

with $\sigma^2(\mu) = v(\mu)^T \Sigma v(\mu)$. *If in addition* $f(k)/p_{\pi_o}(k) \neq c_2$ *and, if* $c_1 > 0$, $f(k)/p_{\pi_o}(k) \neq c_1$ *for any* $k = 0, 1, \ldots, m$, *then*

$$\Pr\left(n^{1/2} \frac{\hat{\pi} - \pi_o}{V(\pi_o)} \leq 0 \right) \to \frac{1}{2} = \Phi(0). \tag{5.3}$$

The proof of this theorem is also given in the Appendix.

The functions $V(\pi_{o+})$ and $V(\pi_{o-})$ for the E-estimator with $c_1 = 0$ and $c_2 = 1$ are visualized in Figure 3 in case of a binomial model with $m = 8$. They can be compared with $V(\pi_o)$ of the MLE which is equivalent to the asymptotic standard deviation of the MLE. Before discussing the implications of Theorem 1, we state an immediate but useful corollary.

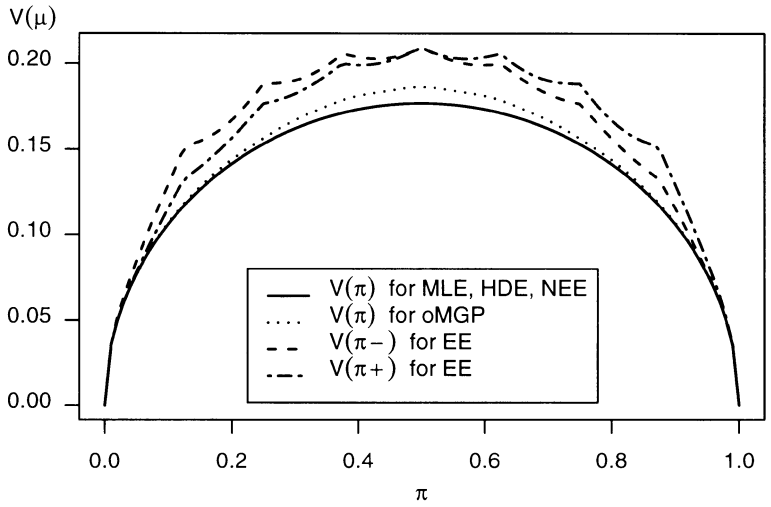

FIG. 3. *Functions* $V(\pi_{o+})$ *and* $V(\pi_{o-})$ *of the E-estimator with* $c_1 = 0$ *and* $c_2 = 1$ (*EE*) *in case of a binomial model with* $m = 8$. *Superimposed is* $V(\pi)$ *of the MLE, HDE, NEE and the oMGP-estimator witb* $b = 0.252$.

COROLLARY 1. *Suppose that* $0 \le c_1 < 1$ *and* $c_2 \ge 1$, π_o *is the unique minimum of* $H(\pi, f)$ *in* $[0, 1]$, *that* π_o *is an interior point of* $[0, 1]$ *and that* H *is convex in a neighborhood of* π_o. *If* $f(k)/p_{\pi_o}(k) \ne c_2$ *and, if* $c_1 > 0$, $f(k)/p_{\pi_o}(k) \ne c_1$ *for any* $k = 0, 1, \ldots, m$, *then*

$$\Pr\left(n^{1/2}\frac{\hat{\pi} - \pi_o}{V(\pi_o)} \le t\right) \to \Phi(t).$$

The interpretation and implications of Theorem 1 and Corollary 1 are now given in the following remarks.

REMARK 1. If there is no contamination ($\gamma = 0$), we have

$$f(k)/p_{\pi_o}(k) = p_{\pi_*}(k)/p_{\pi_*}(k) = 1 \tag{5.4}$$

for all k. This means that:

1. When $c_2 > 1$, we have that $c_1 < f(k)/p_{\pi_o}(k) = 1 < c_2$ so that the weight function w is differentiable at $f(k)/p_{\pi_o}(k)$ and Corollary 1 establishes the central limit theorem. In this case, $\hat{\pi}$ is asymptotically equivalent to the MLE and hence fully efficient at the binomial model.
2. When $c_2 = 1$, the limiting distribution is not Gaussian but can be approximated by the two Gaussian distributions given in the theorem. We cannot apply (5.3) because the proof does not apply when (5.4) holds. The reason for this is that in establishing (B.2), we need to expand $w(f_n/p_{\pi_*})$ about $f_n = f$, a step which is impossible in this case because it involves expansion at the point $1 = c_2$ at which w' does not exist.

REMARK 2. In the case $c_2 = 1$, the influence function (5.2) (which is computed under the binomial model) does not follow the usual relationship to the asymptotic variance under the binomial model because the latter does not exist. However, the influence function retains a clear interpretation in terms of the bias and it remains desirable to use an estimator with low gross error sensitivity such as the E-estimator with $c_2 = 1$. With this goal in mind, E-estimators with $c_1 < 1$ and $c_2 = 1$ or with $c_1 = 1$ and $c_2 > 1$ are of interest from a robustness point of view.

REMARK 3. If there is contamination ($\gamma > 0$), then provided $f(k)/p_{\pi_o}(k) \ne c_2$ and, if $c_1 > 0$, $f(k)/p_{\pi_o}(k) \ne c_1$ for any $k = 0, 1, \ldots, m$, the corollary establishes the central limit theorem. Otherwise, the limiting distribution is not Gaussian but can be approximated by the two Gaussian distributions given in the theorem.

REMARK 4. Notwithstanding the possible lack of a Gaussian approximation to the asymptotic distribution under the model, the theorem always provides a useful basis for approximate inference. If we let $\hat{V}(\pi_o+)$ denote an

estimator of $V(\pi_o+)$ and $\hat{V}(\pi_o-)$ denote an estimator of $V(\pi_o-)$, then an approximate $100(1-\alpha)\%$ central (but not symmetric) confidence interval for π_o is given by

$$\left[\hat{\pi} - n^{-1/2}\hat{V}(\pi_o+)^{1/2}\Phi^{-1}(1-\alpha/2),\ \hat{\pi} + n^{-1/2}\hat{V}(\pi_o-)^{1/2}\Phi^{-1}(1-\alpha/2)\right].$$

On the other hand, it is simpler to assume that the data are contaminated (so the binomial model does not hold) and that $f(k)/p_{\pi_o}(k) \neq c_2$ and, if $c_1 > 0$, $f(k)/p_{\pi_o}(k) \neq c_1$ for any $k = 0, 1, \ldots, m$ because then we obtain the *usual asymptotic Gaussian* confidence intervals.

REMARK 5. The confidence interval in Remark 4 is not necessarily contained within the parameter space. An interval with this property can be obtained by setting a confidence interval on the logistic scale and back-transforming the endpoints.

REMARK 6. One can obtain a Gaussian limit under the binomial model when, for example, $c_2 = 1$ is replaced by $c_2(n) = 1 + n^{-\alpha}a$ for $a > 0$ and $0 \leq \alpha < 0.5$. Intuitively, because the relative frequencies f_n converge to f at $n^{-1/2}$, this choice of $c_2(n)$ enables quantities like f_n/p_{π_*} to converge to one faster than $c_2(n)$ does and so permits us to make the crucial expansions.

Finite-sample calculations show that, when f_n is a finite-sample realisation of a binomial distribution and $c_1 < c_2 = 1$, then $E[\eta(\widehat{\pi}, f_n)] \neq 0$. It follows from (3.3) that the E-estimator is biased under the true model in finite samples. This unexpected feature of the E-estimator can be fixed, at least to first order, by replacing $c_2 = 1$ by $c_2(n) = 1 + n^{-\alpha}a$ for $a > 0$ and $0 \leq \alpha < 1$. A detailed discussion about a proper choice of $c_2(n)$ will be pursued in subsequent work.

6. Discussion. The issues in robustness for the binomial model are different from those in other models because the binomial model is discrete and has bounded parameter and sample spaces. The bounded parameter and sample space impact on the robustness properties of all reasonable estimators makes the MLE robust. While no improvement on the MLE is possible when $m = 1$, better estimators can be constructed when $m > 1$.

"Huberizing" a criterion function is a natural and powerful way to construct robust estimators. We constructed E-estimators by "Huberizing" the relative entropy. These estimators depend on tuning constants $0 \leq c_1 \leq 1$ and $1 \leq c_2 \leq \infty$. The choice $c_2 = 1$ gives improved first order robustness against "gross error" contamination and the choice $c_1 = 1$ gives improved robustness against truncation. Since we must have $c_1 < c_2$, we cannot treat both types of contamination simultaneously. Under the binomial model the asymptotic distribution of the E-estimator is Gaussian for $c_1 < 1 < c_2$, and non-Gaussian for $c_1 < c_2 = 1$. Consequently, for $c_2 = 1$ the asymptotic variance under the binomial model does not exist, so cannot equal the expected

value of the squared influence function. Under "gross error" contamination for $c_1 < c_2 = 1$ the asymptotic distribution of the E-estimator is Gaussian and hence standard Gaussian inference can be made. Formally, if $f_n(k)/p_{\hat{\pi}}(k) \neq 1$ for all $k = 0, \ldots, m$, we can make Gaussian inference while if $f_n(k)/p_{\hat{\pi}}(k) = 1$ for some $k = 0, \ldots, m$, we can apply Theorem 1. This suggests that, in contrast with formal robustness theory, inference is simpler in practice under the contamination model than under the binomial model.

APPENDIX

A. Proof of Lemma 1. Since π_o is an interior point of $[0, 1]$, there is a $\delta > 0$ such that π_o is an interior point of $A_\delta = [\delta, 1 - \delta]$. Notice that standard arguments can be used to show that for n sufficiently large, $\hat{\pi} \in A_\delta$. Then

$$\sup_{\pi \in A_\delta} |H(\pi, f_n) - H(\pi, f)| \to 0 \qquad \text{almost surely.}$$

which implies that $H(\hat{\pi}, f) \to H(\pi_o, f)$ almost surely and the result obtains. □

B. Proof of Theorem 1. As $H(\pi, f)$ is convex in a neighborhood of π_o, the estimating function $\eta(\pi, f)$ is non-decreasing as a function of π in this neighborhood. Then

$$n^{1/2} \frac{\hat{\pi} - \pi_o}{V} \leq t \Leftrightarrow \eta(\pi_n(t),\ f_n) \geq 0,$$

where $\pi_n(t) = \pi_o + n^{-1/2} V\, t$.

Let $Z_n(k) = n^{1/2}(f_n(k) - f(k))$ and note that we can write

$$Z_n = (Z_n(0), \ldots, Z_n(m))^T = n^{-1/2} \sum_{i=1}^{n} (I_i - f),$$

where $I_i = (I(Y_i = 0), \ldots, I(Y_i = m))^T$ are independent multinomial$(1, f)$ random variables. It follows immediately from the central limit theorem that $Z_n \to Z \sim N(0, \Sigma)$.

We can write

$$\begin{aligned} n^{1/2}\eta(\pi_n(t), f_n) &= -\sum_{k=0}^{m} w\left(\frac{f_n(k)}{p_{\pi_n(t)}(k)}\right) s_{\pi_n(t)}(k) \left\{Z_n(k) + n^{1/2} f(k)\right\} \\ &= -\sum_{k=0}^{m} w\left(\frac{f_n(k)}{p_{\pi_n(t)}(k)}\right) s_{\pi_n(t)}(k) Z_n(k) \qquad \text{(B.1)} \\ &\quad -n^{1/2} \sum_{k=0}^{m} w\left(\frac{f_n(k)}{p_{\pi_n(t)}(k)}\right) s_{\pi_n(t)}(k) f(k). \end{aligned}$$

Suppose that $\frac{f(k)}{p_{\pi_o}(k)} = c_2$ and, if $c_1 > 0$, $\frac{f(k)}{p_{\pi_o}(k)} = c_1$. Then $\frac{f(k)}{p_{\pi_n(t)}(k)} \neq c_1, c_2$ because $t \neq 0$ and the derivative of w exists at $\frac{f(k)}{p_{\pi_n(t)}(k)}$. On the other hand, if

$\frac{f(k)}{p_{\pi_o}(k)} \neq c_2$ and, if $c_1 > 0$, $\frac{f(k)}{p_{\pi_o}(k)} \neq c_1$, then there is a neighborhood of $\frac{f(k)}{p_{\pi_o}(k)}$ which excludes c_1, c_2 and for n sufficiently large, with probability tending to one, $\frac{f(k)}{p_{\pi_n(t)}(k)}$ is in this neighborhood so the derivative of w exists at $\frac{f(k)}{p_{\pi_n(t)}(k)}$. In fact, the second derivative also exists at this point. Thus we can expand w about$f_n = f$ to obtain

$$n^{1/2}\eta(\pi_n(t), f_n) = -\sum_{k=0}^{m} w\left(\frac{f(k)}{p_{\pi_n(t)}(k)}\right) s_{\pi_n(t)}(k) Z_n(k) + n^{1/2}\eta(\pi_n(t), f) - \sum_{k=0}^{m} w'\left(\frac{f(k)}{p_{\pi_n(t)}(k)}\right) \frac{f(k)}{p_{\pi_n(t)}(k)} s_{\pi_n(t)}(k) Z_n(k) + o_p(1).$$

Next, applying $w'(x)x = -w(x) + I(c_1 \leq x \leq c_2)$, we obtain

$$n^{1/2}\eta(\pi_n(t), f_n) = -\sum_{k=0}^{m} I\left(c_1 \leq \frac{f(k)}{p_{\pi_n(t)}(k)} \leq c_2\right) s_{\pi_n(t)}(k) Z_n(k) + n^{1/2}\eta(\pi_n(t), f) + o_p(1).$$

Set

$$n^{1/2}\psi(\pi_n(t), Z_n) = -\sum_{k=0}^{m} I\left(c_1 \leq \frac{f(k)}{p_{\pi_n(t)}(k)} \leq c_2\right) s_{\pi_n(t)}(k) Z_n(k) = -n^{-1/2}\sum_{i=1}^{n} v(\pi_n(t))^T (I_i - f)$$

so that

$$n^{1/2}\eta(\pi_n(t), f_n) = n^{1/2}\psi(\pi_n(t), Z_n) + n^{1/2}\eta(\pi_n(t), f) + o_p(1). \tag{B.2}$$

Now suppose that $t > 0$. Then

$$\begin{aligned}
&\Pr\{n^{1/2}\eta(\pi_n(t), f_n) \geq 0\} - \Phi(t) \\
&\quad = 1 - \Pr\{\eta(\pi_n(t), f_n) \leq 0\} - \Phi(t) \\
&\quad = \Phi\left(-n^{1/2}\frac{\eta(\pi_n(t), f)}{\sigma(\pi_o+)}\right) - \Pr\{\eta(\pi_n(t), f_n) \leq 0\} \\
&\qquad + \Phi\left(n^{1/2}\frac{\eta(\pi_n(t), f)}{\sigma(\pi_o+)}\right) - \Phi(t).
\end{aligned}$$

Applying (B.2), we have therefore that

$$\begin{aligned}
&\left|\Pr\{\eta(\pi_n(t), f_n) \geq 0\} - \Phi(t)\right| \\
&\quad \leq \sup_x \left|\Phi(x) - \Pr\left\{n^{1/2}\psi(\pi_n(t), Z_n)/\sigma(\pi_o+) \leq x\right\}\right| \\
&\qquad + \left|\Phi\left(n^{1/2}\frac{\eta(\pi_n(t), f)}{\sigma(\pi_o+)}\right) - \Phi(t)\right|.
\end{aligned} \tag{B.3}$$

Now

$$\mathrm{Var}\left(n^{-1/2}\sum_{i=1}^{n} v(\pi_n(t))^T(I_i - f)\right) = v(\pi_n(t))^T \Sigma v(\pi_n(t)) \to \sigma(\pi_o+)^2$$

and, by Minkowski's inequality,

$$\mathrm{E}\left[\left|v(\pi_n(t))^T(I_i - f)\right|^3\right] \leq \left[\sum_{k=0}^{m}\left\{\left|s_{\pi_n(t)}(k)\right|^3 f(k)(1-f(k))\{(1-f(k))^2 + f(k)^2\}\right\}^{1/3}\right]^3.$$

We can apply the Berry-Esseen theorem to show that

$$\sup_x |\Phi(x) - Pr\{n^{1/2}\psi(\pi_n(t), Z_n)/\sigma(\pi_o+) \leq x\}| \to 0.$$

The same argument applies when $t < 0$ with $\sigma(\pi_o+)$ replaced by $\sigma(\pi_o-)$.

Next, write

$$\begin{aligned} n^{1/2}\eta(\pi_n(t), f) &= n^{1/2}\{\eta(\pi_n(t), f) - \eta(\pi_o, f)\} \\ &= -n^{1/2}\sum_{k=0}^{m}\left\{w\left(\frac{f(k)}{p_{\pi_n(t)}(k)}\right) - w\left(\frac{f(k)}{p_{\pi_o}(k)}\right)\right\} s_{\pi_n(t)}(k) f(k) \\ &\quad -n^{1/2}\sum_{k=0}^{m} w\left(\frac{f(k)}{p_{\pi_o}(k)}\right)\{s_{\pi_n(t)}(k) - s_{\pi_o}(k)\} f(k). \end{aligned}$$

The function $s_\pi(k)$ is differentiable so the second term converges to

$$-t\,V\sum_{k=0}^{m} w\left(\frac{f(k)}{p_{\pi_o}(k)}\right) s'_{\pi_o}(k) f(k).$$

The first term can be written as

$$-t\,V\sum_{k=0}^{m}\left\{\frac{w\left(\frac{f(k)}{p_{\pi_n(t)}(k)}\right) - w\left(\frac{f(k)}{p_{\pi_o}(k)}\right)}{\frac{f(k)}{p_{\pi_n(t)}(k)} - \frac{f(k)}{p_{\pi_o}(k)}}\right\}\left\{\frac{\frac{f(k)}{p_{\pi_n(t)}(k)} - \frac{f(k)}{p_{\pi_o}(k)}}{n^{-1/2}t\,V}\right\} s_{\pi_n(t)}(k) f(k)$$

which converges to

$$\begin{cases} t\,V\sum_{k=0}^{m} w'\left(\frac{f(k)}{p_{\pi_o+}(k)}\right)\frac{f(k)}{p_{\pi_o}(k)}\, s_{\pi_o}(k)^2 f(k), & \text{when } t > 0 \text{ and} \\ t\,V\sum_{k=0}^{m} w'\left(\frac{f(k)}{p_{\pi_o-}(k)}\right)\frac{f(k)}{p_{\pi_o}(k)}\, s_{\pi_o}(k)^2 f(k), & \text{when } t < 0. \end{cases}$$

In order that the second term in (B.4) converges to zero, V must assume the form stated in the theorem.

When $t = 0$, we have $\pi_n(t) = \pi_o$ so (B.2) becomes $n^{1/2}\eta(\pi_o, f_n) = n^{1/2}\psi(\pi_o, Z_n) + o_p(1)$, because $\eta(\pi_o, f) = 0$ by definition of π_o. Provided that $f(k)/p_{\pi_o}(k) \neq c_1, c_2$ for any k, it follows that

$$\Pr(\eta(\pi_o, f_n) \geq 0) \approx \Pr(n^{1/2}\psi(\pi_o, Z_n)/\sigma(\pi_o) \geq 0) \to \Phi(0)$$

which proves the theorem. □

Acknowledgments. We are grateful to the Editor, Associate Editor and referees for their helpful comments.

This work was substantially carried out while both authors were at The Australian National University.

REFERENCES

[1] BIANCO, A. M. and YOHAI, V. J. (1997). Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods* (H. Rieder, ed.) 17–34. Springer, New York.

[2] CARROLL, R. J. and PEDERSON, S. (1993). On robustness in the logistic regression model. *J. Roy. Statist. Soc. Ser. B* **55** 693–706.

[3] CHRISTMANN, A. (1997). High breakdown point estimators in logistic regression. In *Robust Statistics, Data Analysis, and Computer Intensive Methods* (H. Rieder, ed.) 79–90. Springer, New York.

[4] COPAS, J. B. (1988). Binary regression models for contaminated data. *J. Roy. Statist. Soc. Ser. B* **50** 225–265.

[5] CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464.

[6] DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum, and J. J. L. Hodges, ed.) 157–184. Wadsworth, Belmont, CA.

[7] HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. dissertation, Univ. California, Berkeley.

[8] HAMPEL, F. R., RONCHETTI, E. M, ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

[9] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **36** 1753–1758.

[10] 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. *J. Amer. Statist. Assoc.* **84** 460–466.

[11] LINDSAY, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22** 1081–1114.

[12] MARKATOU, M., BASU, A. and LINDSAY, B. (1997). Weighted likelihood estimation equations: The discrete case with applications to logistic regression. *J. Statist. Plann. Inference* **57** 215–232.

[13] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, New York.

[14] PREGIBON, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38** 485–498.

[15] READ, T. R. C. and CRESSIE, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer, New York.

[16] SIMPSON, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82** 802–807.

[17] SIMPSON, D. G., CARROLL, R. J. and RUPPERT, D. (1987). *M*-estimation for discrete data: Asymptotic distribution theory and implications. *Ann. Statist.* **15** 657–669.

[18] STEFANSKI, L. A., CARROLL, R. J. and RUPPERT, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* **2** 413–424.

[19] VICTORIA-FESER, M.-P. and RONCHETTI, E. (1997). Robust estimation for grouped data. *J. Amer. Statist. Assoc.* **92** 333–340.

DEPARTMENT OF PHYSICS
AND MATHEMATICS
ZÜRICH UNIVERSITY OF APPLIED SCIENCES
8401 WINTERTHUR
SWITZERLAND
E-MAIL: Andreas.Ruckstuhl@zhwin.ch

FACULTY OF MATHEMATICAL STUDIES
UNIVERSITY OF SOUTHAMPTON
HIGHFIELD, SOUTHAMPTON
HANTS SO17 1BJ
UNITED KINGDOM
E-MAIL: A.H.Welsh@maths.soton.ac.uk