1. INTRODUCTION

We consider robust estimation in the logistic regression model

\[ \Pr(Y = 1|x) = F(x^t \beta_L), \quad F(v) = \{1 + \exp(-v)\}^{-1}. \]  

(1.1)

Along with the usual logistic model, we will be concerned with a misclassification model in which each response is misclassified with probability \( \gamma \), so that

\[ \Pr(Y = 1|x) = F(x^t \beta_{Mc}) + \gamma \left\{ 1 - 2F(x^t \beta_{Mc}) \right\} = G(x^t \beta_{Mc}, \gamma). \]  

(1.2)

In (1.1), we have used \( \beta_L \) to refer to the true regression parameter under the logistic model, while \( \beta_{Mc} \) is the true regression parameter under the misclassification model.

In an important paper, Copas (1988) contrasts two forms of robust estimates:

- Robust/resistant estimates due to Pregibon (1982);
- The misclassification maximum likelihood estimate (mle) for model (1.2), corrected to be approximately consistent for the logistic model (1.1).

Copas concludes that the latter is a preferable method of estimation at the logistic model, as it seems to have the robustness properties of the former method while at the same time having less small to moderate sample size bias. Model (1.2) also has uses when response misclassification is of interest (Copas, 1988); these will not be discussed in detail here. Copas focused on small values of \( \gamma \) and the use of (1.2) in generating robust estimators and diagnostics. See also Ekholm & Palmgren (1987).

Despite this work, a number of important problems remain to be investigated. Among these are the following:

- Pregibon’s estimate is inconsistent at the logistic model. Thus, it might be the case that the bias observed in Copas’ study is really not so much a function of the entire class of robust/resistant estimates as it is a function of the particular method employed. Our results will indicate that some forms of robust/resistant estimates can have much less small sample bias than previously thought.

- The corrected misclassification mle is only approximately consistent, with no large sample theory available. While a large sample theory can be constructed using the techniques of
Stefanski (1985), instead we will propose and study an estimate which is closely related to the misclassification mle but which is consistent at the logistic model. An advantage of the method compared to other robust estimates is that it has a more easily interpretable tuning constant.

The outline of this paper is as follows. In Section 2, we will give an overview of the robust estimates previously proposed in the literature. Our emphasis is on methods which downweight on the basis of extreme leverage and/or extreme predicted values, the so-called “Mallows”-class (Mallows, 1975; Hampel, et al. 1986, §6.3). In Section 3, we note that estimates derived under Copas’ misclassification model are not consistent at the logistic model. We provide a simple method of correcting the misclassification estimate, which is a member of the Mallows-class. The estimate also retains the highly useful interpretability of misclassification; weights are defined by using a tuning constant which can be interpreted as a nominal misclassification rate.

In Section 4, we illustrate the behavior of these estimates on three sets of data. In Section 5, we investigate the small sample bias properties of the robust/resistant estimates, and compute these biases numerically in the examples.

2. ROBUST ESTIMATES

2.1. Introduction

The primary robust estimates for the logistic model (1.1) solve equations of the form

\[ 0 = \sum_{i=1}^{n} w_i x_i \left\{ Y_i - F(x_i^t \beta) - c(x_i, \beta) \right\}, \quad (2.1) \]

where the \( \{w_i\} \) are weights which may depend on the response. If \( w_i \equiv 1 \) and \( c(x_i, \beta) \equiv 0 \), then (2.1) yields the usual logistic regression estimate. If \( w_i = w(x_i, x_i^t \beta) \) and \( c(x_i, x_i^t \beta) \equiv 0 \), then the weights depend only on the design and we are in the so-called “Mallows” class (Mallows, 1975; Hampel, et al. 1986, §6.3). Finally, if \( w_i = w(x_i, x_i^t \beta, Y_i) \), then we are in the “Schweppe” class, see Künsch, et al. (1989) and Hampel, et al. (1986, §6.3).

Schweppe estimates as written include just about anything, but they only make good sense when the estimating equation is conditionally unbiased given \( x \), i.e.,

\[ 0 = E \left[ w(x, x^t \beta, Y) \left\{ Y - F(x^t \beta) - c(x, \beta) \right\} \right]. \quad (2.2) \]
Pregibon’s estimates, which downweight the deviances by tapering, do not satisfy (2.2) since they set \( c(x, \beta) \equiv 0 \) and, for a tuning constant \( h \),

\[
w(x, x^t \beta, Y) = w(Y - F(x^t \beta)), \quad \text{where} \quad w(v) = \begin{cases} 1 & \text{if } |v| \leq \exp(-h/2) \\ -(1/2)h/\log(1 - |v|) & \text{otherwise.} \end{cases}
\]

One could make the Pregibon estimates consistent by changing them to include a debiasing factor \( c(x, \beta) \), but the optimality theory of Künsch, et al. suggests that the result will be no more efficient, and possibly harder to compute, than the estimate proposed by Künsch, et al.

The deviance tapering method of Pregibon can be implemented not just for logistic regression but for any generalized linear model. Just as in logistic regression, the resulting estimating equations are biased in general, and hence the estimates themselves are inconsistent. Mallows estimates in this general situation are just weighted quasilikelihood estimates, while Künsch, et al. provide consistent estimates within the Schweppe class for GLIM’s in canonical form.

### 2.2. The Mallows Class

In the Mallows–type formulation, the weights do not depend on the response directly, but are instead functions only of the design \( \{x_i\} \) and the parameter \( \beta \). Basically, the idea is that points which have high leverage are “dangerous”, and should be downweighted. Estimators \( \hat{\beta} \) in this class are necessarily consistent, since in this case (2.1) is an unbiased estimating equation. While less efficient than the usual logistic mle, the Mallows class has an easily computed small–sample \( O(1/n) \) bias which can be smaller than that of the mle, especially if there are unusual design points. If superscript \((j)\) with parentheses denotes the \( j^{th} \) derivative, define

\[
V_{nj}(\beta) = n^{-1} \sum_{i=1}^{n} w^{(j)}_i x_i x_i^t F^{(1)}(x_i^t \beta);
\]

\[
T_{nj}(\beta) = V_{nj}(\beta) n^{-1} \sum_{i=1}^{n} w^{(2-j)}_i x_i x_i^t V_{n1}^{-1}(\beta) V_{n2}(\beta) V_{n1}^{-1}(\beta) x_i F^{(j)}(x_i^t \beta).
\]

In the last expression, \( w^{(2-j)}_i \) is the \((2-j)^{th}\) derivative of \( w(u,v) \) with respect to \( v \) and evaluated at \( u = x_i \) and \( v = x_i^t \beta \). By a standard argument, consistent solutions \( \hat{\beta} \) to (2.1) are asymptotically normally distributed:

\[
n^{1/2} \left( \hat{\beta} - \beta_L \right) \approx \text{Normal} \left\{ 0, V_{n1}^{-1}(\beta_L) V_{n2}(\beta_L) V_{n1}^{-1}(\beta_L) \right\}.
\]  

(2.3)
The covariance matrix of $\hat{\beta}$ can be estimated consistently by $n^{-1}V_{n1}^{-1}(\hat{\beta})V_{n2}(\hat{\beta})V_{n1}^{-1}(\hat{\beta})$. There are two important subcases: (i) the weights depend only on the design; and (ii) the weights depend only on the probabilities.

**CASE I: Leverage Downweighting** Suppose that $w(x_i, x_i^*\beta) = w(x_i)$, so that we downweight strictly on the basis of leverage. In this case, $w^{(1)} = T_{n1} = 0$. By an analysis similar to that of Copas (1988), at the logistic model the first order bias is

$$E(\hat{\beta} - \beta_L) \approx -(2n)^{-1}\{T_{n2}(\beta_L) + 2T_{n1}(\beta_L)\}.$$  

For regression through the origin (Copas, 1988, p. 230), the bias is

$$E(\hat{\beta} - \beta_L) \approx -(2n)^{-1}\frac{n^{-1}\sum_{i=1}^{n} w_i x_i^3 F^{(2)}(x_i\beta_L) - \sum_{i=1}^{n} w_i^2 x_i^2 F^{(1)}(x_i\beta_L) - \sum_{i=1}^{n} w_i x_i^2 F^{(1)}(x_i\beta_L)}{\left\{n^{-1}\sum_{i=1}^{n} w_i x_i^2 F^{(1)}(x_i\beta_L)\right\}^2}.$$  

For example, if $n = 50$ and the $x$'s consist of 10 each at $\pm 2$, $\pm 1$, 0, then the bias for the mle is approximately 0.019 at $\beta_L = 0.5$ and approximately 0.048 at $\beta_L = 1$. If instead we give weight of only 0.5 to those observations with $x = \pm 2$, then the corresponding biases for the Mallows estimate are 0.027 and 0.062, respectively. This is not by any means a design with any unusual leverage. In practice, the Mallows weight assigned to the design points $\pm 2$ will be very close to 1.0, and there will be little difference in the bias behavior of the two methods.

While this might at first suggest that the Mallows estimators are generally more biased, this is not the case in general. Of particular concern is the case of bias when there are unusual design points, where the classic robustness theory (Hampel, et al., 1986) suggests that robust methods should have smaller biases. We illustrate this numerically. Replace the previous design with the same one except that 2 points are placed at $\pm 5$, and give weight 0.5 to these unusual points, all other weights being 1.0. The mle has approximate biases 0.036 and 0.060 at $\beta_L = 0.5$ and 0.1, while the Mallows estimate has biases 0.029 and 0.056. A Mallows estimate which gives zero weight to the four most extreme design points has biases 0.021 and 0.053.

The main point of this illustration is that in a design with extreme design points, selective downweighting can lead to less biased estimates when compared to the usual mle. The decrease in bias, coupled with the robustness protection, can be worth the loss of efficiency. In the previous example, a weight of 0.25 assigned to the unusual design points leads to a MSE efficiency of over 80% at the logistic model when compared to the mle.
In his North Carolina Ph.D. thesis, L. Stefanski suggests downweighting on the basis of a robust Mahalanobis distance. The basic idea is that those points which are not near the center of the design space should be downweighted. Any of these robust methods depend on how well one can measure the leverage of an individual point; this can be a nontrivial task. In the leukemia data listed by Cook & Weisberg (1982), the most influential point occurs in a situation where there are three identical design points. As a group, these points are highly leveraged, but automatic methods for measuring leverage, such as that given above, may have problems identifying that each of the points individually has leverage. More details on this method are given in Section 4.

**CASE II: Prediction Downweighting** An alternative is to set \( w(x_i, x_t^i) = w(x_t^i) \), so that extreme fitted probabilities are downweighted. An example of this kind is

\[
  w(x_t^i) = \frac{1}{F(x_t^i)} \frac{1 - F(x_t^i)}{[1 - F(x_t^i)]^{\lambda}},
\]

for constants \((c, \lambda)\). The choices \((c, \lambda) = (0,0)\) or \((1,-1)\) yields the usual logistic estimate. The choice \((c, \lambda) = (1,0)\) weights according to the variance of \(Y\) given \(x\), and will handle outliers at very low or very high predicted probabilities, as does Copas’ misclassification estimate.

### 2.3. Misclassification Estimate

For the misclassification model (1.2), the mle \( \hat{\beta}_{Mc} \) is an M-estimator solving

\[
  0 = \sum_{i=1}^{n} w_i(x_i^t \beta; \gamma) x_i \left\{ Y_i - G(x_i^t \beta; \gamma) \right\},
\]

where

\[
  w_i(x_i^t \beta; \gamma) = (1 - 2\gamma) F(x_i^t \beta) \left[ G(x_i^t \beta; \gamma) \{ 1 - G(x_i^t \beta, \gamma) \} \right]^{-1},
\]

and where \( G \) is defined in (1.2). At the logistic model, \( \hat{\beta}_{Mc} \) converges not to \( \beta_{Mc} \) defined by (1.2), but instead to \( \beta_{Mc^*} \), which is a solution to the equation

\[
  0 = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} w_i(x_i^t \beta_{Mc^*}; \gamma) x_i \left\{ F(x_i^t \beta_{L}) - G(x_i^t \beta_{Mc^*}, \gamma) \right\}.
\]

Recognizing that \( \hat{\beta}_{Mc} \) is inconsistent for \( \beta_L \) at the logistic model, Copas (1988) suggested a bias–corrected version appropriate for small \( \gamma \), see his equation (27). This estimate is inconsistent at the logistic model, although this is of practical concern primarily for larger values of \( \gamma \).

### 2.4. Influence Functions

It can be shown that the Mallows leverage downweighted estimates have bounded influence functions when \( \|x\| w(x) \) is bounded. The misclassification mle has a bounded influence function
only when \( \|x\| F(x^t \beta_{Mc}) \{1 - F(x^t \beta_{Mc})\} \) is bounded, which occurs, in general, only for a scalar predictor \( x \); otherwise a ridge can be formed. Hence, estimation may be unstable for small sample sizes. The Mallows prediction downweighted estimates have a similar behavior: their influence function is bounded when \( \|x\| w(x^t \beta) \) is bounded.

3. CONSISTENT ESTIMATION AT THE LOGISTIC MODEL

Let \( \hat{\beta}_{Mc} \) be the misclassification estimate, solving (2.4). As discussed above, while this estimate is consistent for the parameter \( \beta_{Mc} \) in the misclassification model (1.2), it is inconsistent for \( \beta_L \) in the logistic model (1.1). Copas (1988) constructs a correction to \( \hat{\beta}_{Mc} \) which is approximately consistent at the logistic model when the nominal misclassification rate \( \gamma \) is small, but this estimate is not consistent in general. In this section, we introduce a consistent estimate of \( \beta_L \) which is based on the misclassification estimate \( \hat{\beta}_{Mc} \) and has a easily interpreted tuning constant.

The simplest method is to use the weights from the misclassification estimate, namely those given by \( w(x_i^t \hat{\beta}_{Mc}, \gamma) \), see (2.5). Define \( \hat{\beta} \) as the solution to

\[
0 = \sum_{i=1}^{n} w(x_i^t \hat{\beta}_{Mc}, \gamma) x_i \{ Y_i - F(x_i^t \beta) \}. \tag{3.1}
\]

Equation (3.1) will usually have a unique solution, since it is the estimating equation for a weighted logistic regression with weights \( w(x_i^t \hat{\beta}_{Mc}, \gamma) \). It is easily shown that \( \hat{\beta} \) follows the asymptotic theory (2.3) with \( w_i = w(x_i^t \beta_{Mc*}, \gamma) \), where \( \beta_{Mc*} \) is defined just following (2.5).

The “tuning constant” for the solution to (3.1) is \( \gamma \). The tuning constant can be interpreted as the nominal amount of misclassification assumed in constructing the weights.

Equation (3.1) can also be interpreted as correcting the misclassification estimating equation (2.4) in order to make it unbiased at the logistic model. By subtracting the expectation of (2.4) from (2.4) itself, one obtains (3.1).

It is also possible to replace \( \hat{\beta}_{Mc} \) by Copas’ corrected estimate, with the same theory applying.

The form (3.1) is also instructive as to the behavior of the resulting estimate. From (2.5) we see that the weights become small when \( |x_i^t \beta| \) is large, with \( \gamma \) controlling how extreme the fitted probabilities have to be before significant downweighting occurs. For those \( \gamma \) not near 0, the weights become nearly proportional to \( F(x^t \beta) \{1 - F(x^t \beta)\} \), an option discussed in Section 2.2.

In the strict sense of the term, the solution \( \hat{\beta} \) to (3.1) and the misclassification estimate \( \hat{\beta}_{Mc} \) are not robust, because they have unbounded influence functions. As discussed in Section 2.5, the
potential problem comes when \( \|x\| \) is large but \( |x_i^t\beta| \) is small. See Sections 4.2 and 4.4 for examples which require large values of \( \gamma \) before the estimates are insensitive to the deletion of a small group of points.

4. EXAMPLES

4.1. Introduction

We investigate three examples, one of which involves an extreme leverage and prediction outlier, i.e., an obvious leverage point with extreme predicted probabilities. The second involves an observation which is only an extreme prediction outlier, while the third involves points which are moderate prediction outliers but have considerable leverage. We shall see that the methods behave differently in these data sets.

We first define the Mallows estimates which downweight strictly on the basis of leverage. Let \( p = \text{dim}(x) \). We have found that the following method works reasonably well in practice. Write \( x_i^t = (1, z_i^t)^t \), and let \((\mu, M)\) be a “robust” estimate of the center and covariance matrix of the \( \{z_i\} \). Let \( \psi_{1b} \) be any odd function, and define \( \psi_{2b}(v) = \psi_{1b}^2(v)/\xi \), where \( \xi = E\psi_{1b}^2(\|Z_{p-1}\|) \) and \( Z_{p-1} \) is a \((p-1)\)-dimensional normal random variable with zero mean and identity covariance. Define \( u_{ib}(v) = \psi_{ib}(v)/v \). The estimates \((\mu, M)\) are the solutions to:

\[
\begin{align*}
  n^{-1} \sum_{i=1}^n u_{1b} \left[ \left\{ (z_i - \mu)^t M^{-1} (z_i - \mu) \right\}^{1/2} \right] (z_i - \mu) &= 0; \\
  n^{-1} \sum_{i=1}^n u_{2b} \left[ \left\{ (z_i - \mu)^t M^{-1} (z_i - \mu) \right\}^{1/2} \right] (z_i - \mu) (z_i - \mu)^t &= M.
\end{align*}
\]

In the calculations, we used the trisquared redescending function

\[
\psi_{1b}(v) = v \left\{ 1 - (v/b)^2 \right\}^3 I(|v| \leq b).
\]

If \( d_i = (z - \mu)^t M^{-1} (z - \mu) \), then the Mallows weights to be used in (2.1) are

\[
w(x_i, x_i^t\beta) = w(x_i) = u_{1b} \left[ \left\{ d_i/(p-1) \right\}^{1/2} \right].
\]

Of course, in (2.1) we set \( c(x_i, \beta) \equiv 0 \). Note that these weights can redescend to zero, so that points which are extremely outlying in the design space receive zero weight. Note too that different estimates of center and location can cause the weights to change. For example, the estimates of
$(\mu, M)$ used here do not have the highest possible breakdown point, see Simpson, et al. (1991) for discussion. Ruppert (1991) discusses computation of S-estimates, which are highly robust estimates of $(\mu, M)$.

For the Mallows leverage downweighted estimates, we used the tuning constant $b = 8$, which leads to weights with the following behavior. If the distances from the center are $\{d_i/(p - 1)\}^{1/2}$, $b = 8$ gives weight .75 or greater to those points with distance $\leq 2.5$, weight .50 or greater if the distance is $\leq 3.6$ and weight .25 or greater if the distance is $< 5.0$. For the misclassification mle and its modification defined by (3.1), we used nominal misclassification rates of $\gamma = 0.03$ and $0.01$.

We also computed the behavior of the Schweppe-type estimates of Künsch, et al. (1989). These are defined as follows. For a given matrix $M$, the estimates satisfy (2.1), with

$$w_i = u_{1b}\left\{Y_i - F(x_i^T\beta) - c(x_i^T\beta, x_i^T M^{-1} x_i)((x_i^T M^{-1} x_i)^{1/2})\right\},$$

where the function $c(a, b)$ is chosen so that the right hand side of (2.1) has mean zero when evaluated at $\beta_{Mc}$ and the logistic model, see Künsch, et al. (1989) for formulae. The matrix $M$ can be estimated using equations (2.8) and (2.9) of Künsch, et al. (1989), or, alternatively, it could be a robust scatter matrix estimate.

Consistent covariance matrix estimates for Copas’ corrected method can be constructed using the techniques of Stefanski (1985), and are available in technical report form from the first author.

4.2. Food Stamp Data

These data are discussed in Künsch, et al. (1989, p. 465). There are three predictors of the response $Y = \text{participation in the federal food stamp program}$: (i) tenency; (ii) supplemental income; (iii) log(monthly income + 1). The first two predictors are binary. There is a single case, #5, which is isolated in the design space for variable (iii) and appears to be a response outlier with $Y = 1$. Künsch, et al. (1989) also suggest that case #66 is somewhat outlying. The interesting coefficient is the fourth component $\beta_4$, corresponding to monthly income, which changes considerably in numerical value and significance level with the deletion of case #5 and to a lesser extent with the deletion of case #66.

The last component of $x$ is at least 4.4, except for case #5 where this component was 0. The fitted logistic coefficient without case #5 was (-5.3, 1.8, -7, 1.1). The fitted probabilities were at least .33, except for case #5 which had fitted probability .005. In sum, case #5 is both a leverage
and a prediction outlier, and so is amenable for analysis by all the methods.

We see from Table 1 that the leverage downweighting estimate and the estimate of Künsch, et al. downweight case #5 sufficiently to obtain the desired change from the logistic mle. In this case, we must choose \( \gamma \) rather large, 0.04, in order for the misclassification downweighted and Copas’ correction to the misclassification mle to show the desired effect. For comparison, we show the results for \( \gamma = 0.03 \), which do not pick up the outlier.

### 4.3. Leukemia Data

These data are listed in Cook and Weisberg (1982, page 193). The response is survival for 52 weeks, and the two predictors are white blood cell count and AG status. One observation, #17, appears to be a response outlier, but it is in a group of three points with identical extreme white blood cell values. Without #17, the logistic coefficients were (.2, -23.5, 2.6). Observation #17 had prediction from this fit of less than .001, but the response was \( Y = 1 \). This is an extreme prediction outlier. However, #17 is not an extreme leverage outlier by itself, and quantifying its leverage is a difficult task. The Mallows estimate with misclassification downweighting, Copas’ corrected version of the misclassification mle and Künsch, et al.’s method all show the desired large change from the ordinary logistic mle. Case #17 is a likely misclassified point, under model (1.2), and the new parameter values reflect this. The Mallows leverage downweighting estimate performs poorly here, as it had difficulty assessing the leverage of the unusual observation. There are other ways of assessing leverage, of course, and they might have led to better performance of the these estimates.

### 4.4. Rare Event Data Set

The final data set consists of 300 observations with two predictors \((X_1, X_2)\), generated as follows. First, 296 observations were selected from an ongoing study of diet and breast cancer. The 37 cases with \( Y = 1 \) were included along with a non-random subset of the cases \( Y = 0 \). Here, \( X_j \) ranges from \(-j\) to \( j\), for \( j = 1, 2 \). In these data, the logistic regression coefficients were \((-2.6, 1.5, -0.9)\). The 10\(^{th}\) percentile of the probabilities from this fit was .05, while the 90\(^{th}\) was .24.

We then added 4 artificial outliers to the data, all of which were extreme in the design space. These were \((Y, X_1, X_2) = (1, -25, 4.5), (1, -25, 5.5), (0, -25, -4.5) \) and \((0, -25, -5.5)\). The probabilities from the original logistic fit were .0011, .0005, .85, .94. Thus, the first two observations are extreme in the prediction space. The latter two are not, although they are extreme compared
to the rest of the data.

Our hypothesis was that all the methods would easily handle the first two generated points. However, only the Mallows leverage downweighted and the Künsch, et al. estimates handle the last two points, see Table 3.

4.5. Summary

The key feature of these examples that they indicate that leverage and prediction each have roles to play in the performance of fitting methods. Methods which only handle leverage can get fooled by a leverage group. Methods which only handle extreme predictions can get fooled by leverage points which have unusual predictions not near 0 or 1. The method of Künsch, et al. (1989) works well on all these examples. However, there surely are examples which can fool even this method, and we by no means recommend it uniquely above the others. The behavior of the consistent misclassification estimate defined by (3.1) appears to be at least as good as that of the approximate misclassification estimate suggested by Copas (1988). The asymptotic theory of our estimate has the advantage of being both standard and simple, see (2.3).

5. REMARKS ON SMALL SAMPLE BIAS

5.1. Theoretical Developments

By means of formal asymptotic expansions, one can compute the bias of the robust/resistant estimates discussed in this paper to order $O(n^{-1})$. We have already shown how to do this at the logistic model for the Mallows estimates, see Section 2.2, where we found that unusual design points can make the bias of the Mallows leverage downweighted estimate smaller than that of the mle. Bias expansions for the estimates of Künsch, et al. can be constructed as well, but analytic expressions seem infeasible. In the appendix we provide formulae for constructing their approximate biases, as well as the bias of the consistent estimates derived in Section 3.2. In our examples, we computed the necessary derivatives for the Künsch, et al. method numerically.

5.2. Biases in the Foodstamp Data

In Table 1, we present the approximate biases of the various estimates at the parameter value $\beta^t = (-6.88, 2.02, -0.76, 1.33)$ obtained by deleting cases #5 and #66 and refitting by logistic regression. We see in this table that the Mallows estimates, which performed very well on these data, have biases approximately as large as the usual logistic regression estimates, while the Künsch
method has higher bias, although only on the order of 10% of the estimated standard errors. The consistent misclassification estimate of Section 3.1 also has reasonable bias behavior, with bias increasing with $\gamma$.

Copas’ corrected estimate is asymptotically inconsistent, so that one can speak of an asymptotic bias to describe the difference between the correct logistic parameter and the value estimated. With $\gamma = 0.005$, these asymptotic biases are small, namely $(-.14, .009, -.001, .025)^t$, while with $\gamma = 0.04$, the asymptotic biases are larger, namely $(2.26, -.134, .038, -.391)^t$. Thus in this particular example, $\gamma$ has to remain fairly small in order that the asymptotic bias not become too large. Such asymptotic biases do not arise with the consistent misclassification estimate.

5.3. Biases in the Leukemia Data

In Table 2, we present the approximate biases of the various estimates at the parameter value $\beta^t = (-1.3, -3.2, 2.26)$ obtained at the logistic regression estimate. It seemed to us nonsensical to evaluate bias at the mle having deleted the unusual point, as Table 2 indicates that the standard errors of any of the estimates having deleted this point are extremely large.

The biases of the Mallows and consistent misclassification estimates are similar in this example. The Künsch method has a much larger bias in the second component. There may be some numerical instability here in our numerical calculation of the necessary derivatives.

5.4. Rare Event Data Set

For this data set we used $\beta^t_L = (-2.55, 1.4, 0.88)$. The bias behavior was much like in the food stamp data, although note that in the third component of $\beta_L$, the robust/resistant methods which account for leverage have much smaller bias than does the ordinary logistic mle.

5.5. Summary

While the Künsch method has the best performance in the examples, it has higher $O(1/n)$ bias. The biases were not unacceptably large for the food stamp and rare event data sets, but were unusually large for the leukemia data.

6. DISCUSSION

Essentially all robust estimates have tuning constants. We have shown in Section 3 that a simple modification of Copas’ misclassification estimate is consistent at the logistic model, is a member of the Mallows prediction downweighting class of robust estimates, and its tuning constant is based
on the misclassification estimate's nominal fraction of misclassification. Its influence function is bounded in certain circumstances.

We also investigated a class of estimates based on leverage downweighting. This class forms a useful complement to the prediction downweighting estimates. In the examples, we found that there are cases where prediction downweighting is required and leverage downweighting has relatively poor robustness behavior, and vice-versa. In some instances, prediction downweighting based on the misclassification model will prove useful only for somewhat large values of $\gamma$, the misclassification rate.

Finally, the estimates of Künsch, et al. (1989) had the best overall robustness performance in the examples, but somewhat worse small sample biases.

ACKNOWLEDGEMENT

Carroll’s research was supported by grants from the National Institute of General Medical Sciences and the National Cancer Institute. It was partially completed during visits to the Center for Decision Sciences at Purdue University, which is supported by the National Science Foundation, and to the Epidemiologic Methods Section of the National Cancer Institute. We thank the reviewers for their patience and many helpful comments.

APPENDIX

A.1. Bias Expansions for M–estimates

We assume that all expectations are conditional on $x_i, \ldots, x_n$, and that $E\Psi_i(\beta) = 0$. The M–estimate is defined as the solution to $0 = \sum_1^n \Psi_i(\beta)$. Write the $j^{th}$ component of $\Psi_i$ as $\Psi_{ij}$. Dropping the argument and letting subscript $\beta$'s denote derivatives, define

\begin{align*}
A_n &= n^{-1} \sum_{i=1}^n E\Psi_{i\beta}; \\
B_n &= n^{-1} \sum_{i=1}^n E\Psi_i \Psi_i^t; \\
C_n &= n^{-1} \sum_{i=1}^n E\Psi_{i\beta} A_n^{-1} \Psi_i; \\

D_n &= \text{vec} \left[ n^{-1} \sum_{i=1}^n \text{trace} \left\{ (E\Psi_{i\beta\beta}) A_n^{-1} B_n A_n^{-1} \right\} \right].
\end{align*}

We claim that

\[ E \left( \hat{\beta} - \beta \right) = -n^{-1} A_n^{-1} \left\{ (1/2) D_n - C_n \right\} + o(1). \] (A.1)

Proving (A.1) involves two simple steps. First recall from standard theory that

\[ \left( \hat{\beta} - \beta \right) = -A_n^{-1} n^{-1} \sum_{i=1}^n \Psi_i + O_p \left( n^{-1} \right), \] (A.2)
so that \((\hat{\beta} - \beta)\) has covariance \(n^{-1}A_n^{-1}B_nA_n^{-1}\). By a Taylor series,
\[
0 = n^{-1} \sum_{i=1}^{n} \left[ \Psi_{ij} + \Psi_{ij}^t (\hat{\beta} - \beta) + (1/2) \text{trace} \left\{ \Psi_{ij\beta} (\hat{\beta} - \beta) (\hat{\beta} - \beta)^t \right\} \right] + O_p(n^{-3/2}).
\]
If \(a_{nj}^t\) is the \(j^{th}\) row of \(A_n\), using (A.2) we thus have
\[
0 = n^{-1} \sum_{i=1}^{n} \Psi_{ij} + a_{nj}^t (\hat{\beta} - \beta) - n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \left\{ \Psi_{ij\beta} - E(\Psi_{ij\beta}) \right\}^t A_n^{-1} \Psi_i \right]
+ (1/2)n^{-1} \sum_{i=1}^{n} \text{trace} \left\{ E(\Psi_{ij\beta}) (\hat{\beta} - \beta) (\hat{\beta} - \beta)^t \right\} + O_p(n^{-3/2}).
\]
Taking expectations yields the result.
It is possible to eliminate the second partial derivatives, as follows. Write \(\Psi_{ij} = \Psi_{ij}(y_i|x_i, \beta)\) and \(h(y|x) = F(x^t \beta)^y \{1 - F(x^t \beta)\}^{1-y}\). By definition, for every \(\beta\),
\[
0 = \sum_{y=0}^{1} \Psi_{ij}(y|x, \beta)h(y|x, \beta).
\]
Taking two derivatives, we obtain
\[
E\Psi_{ij\beta} = -\sum_{y=0}^{1} \left\{ \Psi_{ij\beta}(y)h_{\beta}(y) + h_{\beta}(y)\Psi_{ij}^t(y) + \Psi_{ij}(y)h_{\beta\beta}(y) \right\}.
\]

A.2. Bias Expansion for the Consistent Misclassification Estimate

In this section we compute a bias expansion for the consistent misclassification estimate defined in Section 3.2. Note that \(\hat{\beta}\) can be rewritten as the solution to
\[
0 = \sum_{i=1}^{n} K_i(\hat{\beta}, \gamma, \hat{\beta}_{Mc}) = \sum_{i=1}^{n} w_i(x_i^t \hat{\beta}_{Mc}, \gamma) x_i \left\{ F(x_i^t \hat{\beta}) - G(x_i^t \hat{\beta}_{Mc}, \gamma) \right\}.
\]
Let \(\beta_{Mc}\) be that value for which the expectation of (2.4) is zero under the logistic model, i.e.,
\[
0 = \sum_{i=1}^{n} K_i(\beta_L, \gamma, \beta_{Mc}) = \sum_{i=1}^{n} w_i(x_i^t \beta_{Mc}, \gamma) x_i \left\{ F(x_i^t \beta_L) - G(x_i^t \beta_{Mc}, \gamma) \right\}.
\]
If \(A_{1n} = n^{-1} \sum_1^n (\partial/\partial \beta_{Mc}) K_i(\beta_L, \gamma, \beta_{Mc})\) and \(A_{2n} = n^{-1} \sum_1^n (\partial/\partial \beta_L) K_i(\beta_L, \gamma, \beta_{Mc})\), then by Taylor series arguments,
\[
(\hat{\beta} - \beta_L) \approx -A_{2n}^{-1} A_{1n}(\hat{\beta}_{Mc} - \beta_{Mc}).
\]
If we write \(w_i = w_i(x_i^t \beta_{Mc}, \gamma)\), \(G_i = G(x_i^t \beta_{Mc}, \gamma)\), \(U_i = w_i G_i\) and \(F_i = F(x_i^t \beta_L)\), then by a Taylor series of (4.1) we obtain
\[
0 = n^{-1} \sum_{i=1}^{n} x_i x_i^t \left\{ \left( w_i^{(1)} F_i - U_i^{(1)} \right) \left( \beta_{Mc} - \beta_{Mc} \right) + w_i F_i^{(1)} \left( \hat{\beta} - \beta_L \right) \right\}
+ (1/2)n^{-1} \sum_{i=1}^{n} x_i \left[ \left( w_i^{(2)} F_i - U_i^{(2)} \right) \left\{ x_i^t \left( \hat{\beta}_{Mc} - \beta_{Mc} \right) \right\}^2 + w_i F_i^{(2)} \left\{ x_i^t \left( \hat{\beta} - \beta_L \right) \right\}^2 \right]
+ (1/2)n^{-1} \sum_{i=1}^{n} x_i w_i^{(1)} F_i^{(1)} \left\{ x_i^t \left( \hat{\beta} - \beta_L \right) \left( \beta_{Mc} - \beta_{Mc} \right)^t x_i + x_i^t \left( \beta_{Mc} - \beta_{Mc} \right) \left( \hat{\beta} - \beta_L \right) x_i \right\}.
\]
From (4.2), replace \( (\hat{\beta} - \beta_L) \) by \( -A_{2n}^{-1}A_{1n} (\hat{\beta}_{Mc} - \beta_{Mn}) \) in the last two sums, and recall from standard M-estimator theory that the covariance matrix of \( (\hat{\beta}_{Mc} - \beta_{Mn}) \) under the logistic model is \( n^{-1}V \), where \( V = S_1^{-1}S_4S_1^{-1} \), where

\[
S_1 = n^{-1} \sum_{i=1}^{n} x_i x_i' \left( w_i^{(1)} F_i - U_i^{(1)} \right); \quad S_4 = n^{-1} \sum_{i=1}^{n} x_i x_i' w_i^2 F_i^{(1)}. \]

Taking expectations, we obtain

\[
E \left( \hat{\beta} - \beta_L \right) = -L_1^{-1}L_2 E \left( \hat{\beta}_{Mc} - \beta_{Mn} \right) - (2n)^{-1}L_1^{-1}(L_3 + L_4 + L_5) + o(1), \quad (A.3)
\]

where

\[
L_1 = n^{-1} \sum_{i=1}^{n} x_i x_i' w_i F_i^{(1)}; \quad L_2 = n^{-1} \sum_{i=1}^{n} x_i x_i' (w_i^{(1)} F_i - U_i^{(1)});
\]

\[
L_3 = n^{-1} \sum_{i=1}^{n} x_i w_i F_i^{(2)} a_{3i}; \quad L_4 = n^{-1} \sum_{i=1}^{n} x_i a_{4i} (w_i^{(2)} F_i - U_i^{(2)});
\]

\[
L_5 = n^{-1} \sum_{i=1}^{n} x_i w_i^{(1)} F_i^{(1)} a_{5i}; \quad a_{3i} = x_i' A_{2n}^{-1} A_{1n} V A_{1n} A_{2n}^{-1} x_i;
\]

\[
a_{4i} = x_i' V x_i; \quad a_{5i} = -x_i' A_{2n}^{-1} A_{1n} V x_i - x_i' V A_{1n} A_{2n}^{-1} x_i.
\]

From (A.3) we must compute the first order bias for the misclassification estimate. This is an M-estimate, and the previous section can be used. Detailed calculations show that

\[
E \left( \hat{\beta}_{Mc} - \beta_{Mn} \right) = -(2n)^{-1}S_1^{-1}S_2 + n^{-1}S_1^{-1}S_3, \quad \text{where}
\]

\[
S_2 = n^{-1} \sum_{i=1}^{n} x_i x_i' V x_i \left( w_i^{(2)} F_i - U_i^{(2)} \right); \quad S_3 = n^{-1} \sum_{i=1}^{n} x_i w_i^{(1)} F_i^{(1)} x_i S_1^{-1} x_i.
\]

REFERENCES


ON ROBUSTNESS IN THE LOGISTIC REGRESSION MODEL

R. J. Carroll
Department of Statistics
Texas A&M University
College Station, TX 77843

Shane Pederson
Statistics Group
Los Alamos National Laboratory
Los Alamos, NM 87545

SUMMARY

We investigate robustness in the logistic regression model. Copas (1988) studied two forms of robust estimators, a robust/resistant estimate of Pregibon (1982) and an estimate based on a misclassification model. He concluded that robust/resistant estimates are much more biased in small samples than the usual logistic estimate, and recommends a bias-corrected version of the misclassification estimate. We show that there are other versions of robust/resistant estimates which have bias often approximately the same as and sometimes even less than the logistic estimate; these estimates belong to the Mallows class. In addition, the corrected misclassification estimate is inconsistent at the logistic model; we develop a simple consistent modification. The modified estimate is a member of the Mallows class but unlike most robust estimates, it has an interpretable tuning constant. The results are illustrated on data sets featuring different kinds of outliers.

Keywords: ASYMPTOTICS; BINARY REGRESSION; CONSISTENT ESTIMATION; LEVERAGE; MALLOWS ESTIMATORS; MISCLASSIFICATION; SMALL SAMPLE BIAS EXPANSIONS; OUTLIERS
TABLE 1 – FOODSTAMP DATA SET

For the various methods, parameter estimates are listed, along with standard errors and small sample biases where appropriate. Here, “Consistent” refers to the new estimated defined by (3.1), while “Corrected” refers to the approximate estimate of Copas (1988).

<table>
<thead>
<tr>
<th>Method</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
<th>( \beta_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic mle</td>
<td>-.93</td>
<td>1.85</td>
<td>-.90</td>
<td>.33</td>
</tr>
<tr>
<td>Standard error</td>
<td>(1.62)</td>
<td>(.53)</td>
<td>(.50)</td>
<td>(.27)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.487</td>
<td>.127</td>
<td>-.027</td>
<td>.091</td>
</tr>
<tr>
<td>Mallows Leverage, b = 8</td>
<td>-5.22</td>
<td>1.80</td>
<td>-.66</td>
<td>1.05</td>
</tr>
<tr>
<td>Standard error</td>
<td>(2.63)</td>
<td>(.54)</td>
<td>(.52)</td>
<td>(.44)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.463</td>
<td>.123</td>
<td>-.027</td>
<td>.086</td>
</tr>
<tr>
<td>Künsch, b = 18</td>
<td>-5.22</td>
<td>1.67</td>
<td>-.69</td>
<td>1.04</td>
</tr>
<tr>
<td>Standard error</td>
<td>(2.46)</td>
<td>(.50)</td>
<td>(.50)</td>
<td>(.42)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.634</td>
<td>.197</td>
<td>-.043</td>
<td>.119</td>
</tr>
<tr>
<td>Corrected, ( \gamma = 0.01 )</td>
<td>-.88</td>
<td>1.83</td>
<td>-.89</td>
<td>.32</td>
</tr>
<tr>
<td>Standard error</td>
<td>(1.56)</td>
<td>(.53)</td>
<td>(.50)</td>
<td>(.27)</td>
</tr>
<tr>
<td>Corrected, ( \gamma = 0.03 )</td>
<td>-.71</td>
<td>1.83</td>
<td>-.84</td>
<td>.29</td>
</tr>
<tr>
<td>Standard error</td>
<td>(1.51)</td>
<td>(.58)</td>
<td>(.50)</td>
<td>(.26)</td>
</tr>
<tr>
<td>Corrected, ( \gamma = 0.04 )</td>
<td>-7.90</td>
<td>1.85</td>
<td>-.62</td>
<td>1.51</td>
</tr>
<tr>
<td>Standard error</td>
<td>(4.03)</td>
<td>(.63)</td>
<td>(.55)</td>
<td>(.72)</td>
</tr>
<tr>
<td>Consistent, ( \gamma = 0.03 )</td>
<td>-.81</td>
<td>1.79</td>
<td>-.86</td>
<td>.31</td>
</tr>
<tr>
<td>Standard error</td>
<td>(1.47)</td>
<td>(.40)</td>
<td>(.46)</td>
<td>(.25)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.638</td>
<td>.179</td>
<td>-.045</td>
<td>.119</td>
</tr>
<tr>
<td>Consistent, ( \gamma = 0.04 )</td>
<td>-6.43</td>
<td>1.86</td>
<td>-.71</td>
<td>1.26</td>
</tr>
<tr>
<td>Standard error</td>
<td>(2.54)</td>
<td>(.41)</td>
<td>(.48)</td>
<td>(.42)</td>
</tr>
<tr>
<td>Bias</td>
<td>-6.96</td>
<td>.202</td>
<td>-.052</td>
<td>.130</td>
</tr>
</tbody>
</table>
TABLE 2 – LEUKEMIA DATA SET
For the various methods, parameter estimates are listed, along with standard errors and small sample biases.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic mle</td>
<td>-1.31</td>
<td>-3.18</td>
<td>2.26</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.81)</td>
<td>(1.86)</td>
<td>(.95)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.103</td>
<td>-.976</td>
<td>.270</td>
</tr>
<tr>
<td>Mallows Leverage, $b = 8$</td>
<td>-1.20</td>
<td>-4.05</td>
<td>2.24</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.83)</td>
<td>(2.35)</td>
<td>(.97)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.107</td>
<td>-.934</td>
<td>.269</td>
</tr>
<tr>
<td>Künsch, $b = 18$</td>
<td>.02</td>
<td>-18.55</td>
<td>2.38</td>
</tr>
<tr>
<td>Standard error</td>
<td>(1.08)</td>
<td>(13.25)</td>
<td>(1.17)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.039</td>
<td>-1.97</td>
<td>.317</td>
</tr>
<tr>
<td>Corrected, $\gamma = 0.01$</td>
<td>.22</td>
<td>-22.97</td>
<td>2.51</td>
</tr>
<tr>
<td>Standard error</td>
<td>(1.34)</td>
<td>(26.50)</td>
<td>(1.62)</td>
</tr>
<tr>
<td>Consistent, $\gamma = 0.01$</td>
<td>.19</td>
<td>-20.27</td>
<td>2.42</td>
</tr>
<tr>
<td>Standard error</td>
<td>(1.05)</td>
<td>(12.40)</td>
<td>(1.18)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.121</td>
<td>-1.00</td>
<td>.279</td>
</tr>
</tbody>
</table>
TABLE 3 – SMALL RATE DATA SET

For the various methods, parameter estimates are listed, along with standard errors and small sample biases.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic mle</td>
<td>-2.34</td>
<td>1.05</td>
<td>.07</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.30)</td>
<td>(.54)</td>
<td>(.23)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.052</td>
<td>.041</td>
<td>-.045</td>
</tr>
<tr>
<td>Mallows Leverage, $b = 8$</td>
<td>-2.61</td>
<td>1.49</td>
<td>-.96</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.34)</td>
<td>(.58)</td>
<td>(.33)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.052</td>
<td>.042</td>
<td>-.013</td>
</tr>
<tr>
<td>Künsch, $b = 18$</td>
<td>-2.44</td>
<td>1.22</td>
<td>-.85</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.33)</td>
<td>(.58)</td>
<td>(.32)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.070</td>
<td>.059</td>
<td>-.015</td>
</tr>
<tr>
<td>Corrected, $\gamma = 0.01$</td>
<td>-2.33</td>
<td>1.04</td>
<td>.05</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.30)</td>
<td>(.54)</td>
<td>(.23)</td>
</tr>
<tr>
<td>Corrected, $\gamma = 0.03$</td>
<td>-2.58</td>
<td>1.41</td>
<td>-.35</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.41)</td>
<td>(.66)</td>
<td>(.15)</td>
</tr>
<tr>
<td>Consistent, $\gamma = 0.03$</td>
<td>-2.47</td>
<td>1.27</td>
<td>-.27</td>
</tr>
<tr>
<td>Standard error</td>
<td>(.23)</td>
<td>(.46)</td>
<td>(.20)</td>
</tr>
<tr>
<td>Bias</td>
<td>-.064</td>
<td>.054</td>
<td>-.051</td>
</tr>
</tbody>
</table>