ODDS RATIO ESTIMATION IN THE PRESENCE OF COMPLETE OR QUASI-COMPLETE SEPARATION IN DATA

Rosa Giaimo
Dipartimento di Contabilità Nazionale e Analisi dei Processi Sociali, Università di Palermo, Italia – E-mail giaimo@unipa.it

Domenica Matranga
Dipartimento di Biotecnologie Mediche e Medicina Legale - Sezione Scienze Radiologiche, Università di Palermo, Italia – E-mail domenicamatranga@unipa.it

Giuseppina Campisi
Dipartimento di Scienze Stomatologiche, Università di Palermo, Italia – E-mail campisi@odonto.unipa.it

Abstract

In presence of completely or quasi-completely separated data, the maximum likelihood estimates for the logistic regression parameters do not exist. In medical research the question is of great importance because of the need to obtain finite odds ratios. Statistical packages do not solve the estimation problem with non-overlapped dataset. We suggest to apply the hidden logistic regression model and the MEL estimator of Rousseeuw and Christmann (2003) where a unique solution is graphically obtained by the inspection of the ridge trace of regression parameters (IRT). Alternatively, we introduce a Cross Validation (CV) based method to choose the regularization parameter. A real data-set on oral candidosis affection is considered. Our analysis points out that CV rather than IRT leads to ML estimates with minimum misclassification error rate.

Keywords: Data Separation; Hidden regression; Regularization parameter

1. INTRODUCTION

In several multi-etioologic diseases, logistic regression models are widely used to investigate the relation between a binary response variable, with categories corresponding to presence/absence of a certain disease, and one or more risk factors. The existence, finiteness and uniqueness of maximum likelihood estimates for the logistic regression model depend on the pattern of data points in the observed
space. If data are completely or quasi-completely separated, convergence of the logistic regression model may not occur and odds ratios are not finite.

The best way to solve the problem is to prevent it by employing a well designed study, possibly planned by an expert biostatistician. It often happens in clinical practice that cases and controls are collected without expert guidelines and data successively submitted to statistical analysis. One solution may be to eliminate variables responsible for data separation or to use some method of fuzzy cluster analysis (Giaimo, Matranga, 1997). Such variables, however, are difficult to identify due to their interactions and simultaneous effects. Moreover, despite the complete or quasi-complete separation in the observed sample, clinicians need to understand associations between a specific pathology and its possible risk factors, which are synthesized by finite odds ratios. Thus, in medical research it is not possible to solve the problem by eliminating the variables responsible for data separation.

Another solution regards the possible causes of data separation. The lack of overlapping in data may be caused by misclassification, which can occur in several phases of an observational study. As a consequence, in literature on the subject the problem of data separation has been solved by modifying the maximum likelihood method to correct misclassification in data.

The hidden logistic regression framework, proposed by Ekholm and Palmgren (1982), successively considered by Copas (1988) and more recently by Rousseeuw and Christmann (2003), is based on the hypothesis that the observed response is strongly related but not equal to the unobservable or true latent response. Also Hausman (1998) suggests the maximization of a log-likelihood function of data corrected for misclassification, but model parameters are not separately identified in the linear probability model.

In this paper, we embrace the MEL Estimator of Rousseeuw and Christmann (2003) with misclassification error rate properly chosen. While the method works well when applied to overlapped data, it gives doubtful results in the case of separated data as the choice of the misclassification error parameter can be of great impact, thus giving rise to inconsistent odds ratio estimates.

Whereas the authors suggest choosing the misclassification error parameter through the inspection of the ridge trace (IRT) method, in this paper we propose to implement the MEL Estimator by introducing a Cross Validation (CV) based criterion. Comparison with the IRT method is based on the minimum classification error rate criterion.

Methods are applied to the results of a case-control study designed in a secondary epidemiological hospital base at the Sector of Oral Medicine (University
The paper is organized as follows: in Section 2, the question of data configurations is described and methods to detect separation in data are shown. In Section 3 the MEL Estimator is summarized, while in Section 4 we consider the methodological question of the choice of regularization parameter, which can be obtained either by means of the CV or the IRT method. A comparison is made in Section 5: quasi-complete separation of the OC affection data-set is detected and passed over using the MEL estimation method with regularization parameter computed following the two approaches. Based on our results, the better method seems to be the CV method, which gives rise to the minimum classification error rate.

2. SEPARATION IN DATA

When fitting a logistic regression model to a data-set, characterized by a collection of covariates which completely separates the outcome groups, numerical problems can occur. For example, suppose that the sex of every subject with the outcome present is female and the sex of every subject with the outcome absent is male. Thus, if we know the sex of a person, we certainly know the value of the response variable. In this situation there is no overlap in the distribution of the covariates between the two outcome groups and the data-set is said to have the property of monotone likelihood (Bryson et al., 1981).

Albert and Anderson (1984) show that the maximum likelihood estimate exists if and only if the data set has overlap. Geometrically, it means that the estimate exists if and only if there is no hyperplane which separates successes and failures, where the hyperplane itself can contain both successes and failures (Christmann and Rousseeuw, 2001).

A data-set is completely separated if there exists a vector $\mathbf{b}$ that correctly allocates all observations to their response group, i.e.:

$$b^T x_i > 0 \quad \text{if} \quad y_i = 1$$
$$b^T x_i < 0 \quad \text{if} \quad y_i = 0$$

for $i = 1, \ldots, n$ (1)

The maximum likelihood estimate of $\beta$ does not exist, the loglikelihood goes to zero as iteration progresses and the dispersion matrix becomes unbounded (So, 1993).

If the vector $\mathbf{b}$ is as follows:
\[ b^T x_i \geq 0 \quad \text{if} \quad y_i = 1 \]
\[ b^T x_i \leq 0 \quad \text{if} \quad y_i = 0 \]

for \( i = 1, ..., n \) and there exists at least one \( k \in \{1, 2, ..., n\} \) such that \( b^T x_k = 0 \), the data-set is called quasi-completely separated. The maximum likelihood estimate does not exist, the loglikelihood reduces to a non-zero constant and the dispersion matrix becomes unbounded. When there is no separation, the data present overlaps and maximum likelihood estimates exist and are unique. The following example illustrates the above configurations (see Fig. 1).

Fig. 1: The three mutually exclusive and exhaustive data configurations.

We can observe that vector \( \mathbf{b} = (-5, -2, 1) \) identifies the line \( y = 2x + 5 \) which completely separates the first data-set. The second data-set presents two points lying on the line, so that it is quasi-completely separated. Finally, the third data-set is overlapped by points of the two groups on the same side of the line. It is not possible to draw a straight line separating two groups of an overlapped data-set.

The first problem encountered when performing logistic regression regards the detection of separation in data. Statistical packages, widely used by clinicians and epidemiologists, do not solve the question of odds ratio estimation when there
is complete or quasi complete separation in data points. Rather, they give different outputs.

A recent review on the use of the most popular statistical software programs reveals SAS software to be more accurate than SPSS software in notifying data separation (Webb, 2004). This is because SAS algorithms used to estimate logistic models iterate more than SPSS and, in the case of data separation, provide higher standard errors.

Another approach used to detect data separation has been suggested by Christmann and Rousseeuw (2001). They show that regression depth (Rousseeuw and Hubert, 1999) can be used to measure the amount of separation between successes and failures in data sets that are commonly fitted by binary regression models. Computation time depends on the number of tied observations and on the number of sub-samples (see ref. for details).

We used the SAS empirical approach to recognize data configurations (So, 1993; Stokes et al. 1995). The idea behind this approach is that any convergence method used to maximize the log likelihood must yield a solution giving complete separation, if such a solution exists. In maximizing the log likelihood, there is no checking for complete or quasi-complete separation if convergence is attained in eight or fewer iterations.

Subsequent to the eighth iteration, the probability of the observed response is computed for each observation. If the probability of the observed response is one for all observations, there is a complete separation of data points and the iteration process is stopped.

If the complete separation of data has not been determined and an observation is identified to have a probability larger than 0.95 of the observed response, there are two possible situations. First, there is overlap in the data set, and the observation is an atypical observation of its own group. The iterative process, if allowed to continue, will stop when a maximum is reached. Secondly, there is quasi-complete separation in the data set, and the asymptotic dispersion matrix is unbounded. If any of the diagonal elements of the dispersion matrix for the standardized observation vectors (all explanatory variables standardized to zero mean and unit variance) exceeds 5000, quasi-complete separation is declared and the iterative process is stopped. If either complete separation or quasi-complete separation is detected, a warning message is displayed in the procedure output (SAS Institute Inc., 1999).

3. HIDDEN REGRESSION MODELS

In order to solve the question of parameter estimation in the presence of data
separation, the *hidden logistic regression model* can be considered. This model is an extension of the classical logistic regression model where it is assumed that the real binary response cannot be observed due to the presence of an additional random variable determined by misclassification errors or data-entry errors.

The classical logistic regression model assumes \( n \) binary sample units \( y_i = \pi_i + e_i \) with \( \pi_i = \Pr(Y_i = 1) = f(x_i, \beta) = \left\{1 + \exp\left[-(x_i \beta)\right]\right\}^{-1} \) and independent stochastic errors \( e_i \) with Bernoullian distribution \( \text{bin}(1, \pi_i) \). In the classical meaning, no errors are committed when observing the response. Therefore, if we denote with \( T \) a latent dichotomous variable assuming two possible realizations, success (s) and failure (f), we have:

\[
\Pr(Y_i = 1|T_i = s) = 1 \text{ and } \Pr(Y_i = 1|T_i = f) = 0 \tag{3}
\]

In the hidden version of logistic regression, the above probabilities are \( \delta_1 \) and \( \delta_0 \), respectively, with the reasonable assumption \( 0 < \delta_0 < 0.50 < \delta_1 < 1 \). As an effect, \( \Pr(Y_i = 0|T_i = s) = 1 - \delta_1 \) and \( \Pr(Y_i = 0|T_i = f) = 1 - \delta_0 \) with misclassification errors given by \( \delta_0 \), the probability that a zero is misclassified as a one, and \( 1-\delta_1 \), the probability that a one is classified as a zero (Rousseeuw and Christmann, 2003, p.317).

Based on these misclassification errors, *pseudo-observations* \( \tilde{y}_i \) can be computed (Rousseeuw and Christmann, 2003, pp. 318-319) as the weighted average of \( \delta_0 \) and \( \delta_1 \), with weights given by \( 1-y_i \) and \( y_i \):

\[
\tilde{y}_i = (1 - y_i)\delta_0 + y_i\delta_1 \tag{4}
\]

In a model with \( n \) observations, we obtain \( n \) pseudo-observations \( \tilde{y}_i \), which can be interpreted as the success probability conditional to the most likely estimate of the true status \( t_i \).

The log-likelihood \( L(\beta) \) computed on pseudo-observations (4) is:

\[
L(\beta) = \ln\left(l(\beta)\right) = \sum_{i=1}^{n} \left\{ \tilde{y}_i \ln\left[\pi(x_i)\right] + (1 - \tilde{y}_i)\ln\left[1 - \pi(x_i)\right] \right\} \tag{5}
\]

and the Maximum Estimated Likelihood (MEL) estimator is the value of \( \beta \) which maximizes it.

It is thus demonstrated that the MEL estimator always exists and is unique. As a consequence, the resulting odds ratios \( OR_j = \exp(\beta_j) \) always exist and they are never zero or divergent to infinite. Regarding the uniqueness of \( \beta_j \) and therefore \( OR_j \), it must be stressed that the MEL estimator is unique, given the choice of \( \delta_0 \) and \( \delta_1 \).
The choice of $\delta_0$ and $\delta_1$ can be a matter of great importance, especially if data does not overlap. In fact, in this case, the smaller misclassification errors are the more the MEL estimator tends towards the ML estimator, which itself tends to infinity. Therefore, it is necessary to fix a criterion for the choice of $\delta_0$ and $\delta_1$ in case of non-overlapped data-sets.

4. CHOICE OF THE REGULARIZATION PARAMETER

The question of the choice of misclassification errors does not exist if $\delta_0$ and $\delta_1$ are known \textit{a priori}. Alternatively, if they are not known, we can distinguish between the \textit{symmetric approach} (Copas, 1988) which considers a sufficiently small parameter $\lambda$ (i.e. 0.01 or 0.02) and puts $\delta_0 = \lambda$ and $\delta_1 = 1 - \lambda$, and the \textit{asymmetric approach} of Rousseeuw and Christmann (2003), which takes account of how many 0s and 1s there are in the data-set and leads to the following formulae (Rousseeuw and Christmann, 2003, p. 321-322):

$$
\delta_0 = \frac{\hat{\pi} \delta}{\delta + 1}, \\
\delta_1 = \frac{\hat{\pi} \delta + 1}{\delta + 1}
$$

where $\delta = 0.01$ for default and the average $\hat{\pi} = (1 - \hat{\pi}) \delta_0 + \hat{\pi} \delta_1$ of the pseudo-observations (4) is computed as follows:

$$
\hat{\pi} = \max(\delta, \min(1 - \delta, \sum y_i / n))
$$

For real data-sets with no overlap, the default choice $\delta = 0.01$ cannot be the best choice because it may result too small and cause divergent odds ratios.

In this case, Rousseeuw and Christmann suggest exploring the ridge trace of $\hat{B}(\delta)$ that is to say the plot-set of $\hat{B}_j(\delta)$ as a function of $\delta$. If the functions show a trend rapidly changing from $\delta = \varepsilon > 0$ to greater values of $\delta$, then one would choose the value of $\delta$ when the functions begin to stabilize.

The proposal of this paper is to choose the regularization parameter by means of model selection methods: the underlying idea of such methods regards the optimization of several models over a training set followed by the choice of the model that can optimize a selection criterion, such as a direct estimate of the model’s generalization error or the sum of an error measure and a complexity penalty.

Cross validation (CV) is a method used to estimate generalization error based
on re-sampling. It is often used to choose from amongst various models. There are different CV methods (Hjorth 1994). In particular, we used the Hold out set CV (HOS-CV) and some multifold CV methods.

In HOS-CV, a subset of the available sample units are randomly set aside to serve as a validation set and the remaining samples are used for model selection. This method is often sufficient when abundant data is available, but if we have a sparse quantity of data it is not advisable to set aside a portion of the data-set for testing. Moreover, since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if it happens to get an unlucky split.

The limitations of the holdout can be overcome by repeating the HOS-CV a number of times and averaging the results. We refer to this method as multifold CV.

Within the class of multifold methods (Plutowski, 1996) we considered, in particular, the Disjoint set “Leave-Many-Out” CV (LMO-CV). This method is also known as K-fold-CV because it creates a K-fold partition of the dataset: for each of K experiments, use K-1 folds for training and a different fold for testing. With a large number of folds, the bias of the true error rate estimator will be small, but the variance and the computational time will be very large. On the contrary, with a small number of folds, the bias will be very large but the variance and the computational time will be reduced. In practice, the choice of the number of folds depends on the size of the dataset. For large datasets, 3-fold CV will be accurate, while in general K=10 is required (Breiman 1996; Breiman and Spector 1992).

For small datasets, “Leave-one-out” (LOO-CV), which is the degenerate case of K-fold CV, is better: for a dataset with N observations, perform N experiments; for each experiment use N-1 samples for training and the remaining sample for testing.

LOO-CV is an appealing alternative for many practitioners because it does not require a choice for K, it limits the number of splits to a total of N and it results in the maximum number of samples (N-1) allotted to each training set. However, it is computationally expensive, with the computational demand proportional to N(N-1).

The fundamental principle of CV, in a few words, is that it can predict the observed values from the remaining observations and choose δ which yields the best predictions. Let us consider, for instance, the LOO-CV. If $\hat{y}(x_i)$ denotes the prediction of $y(x_i)$ when the i-th sample unit is removed from data, one CV criterion can be to minimize the following error rate measure:

$$MSE_{cv}(\delta) = \frac{1}{n} \sum_{i=1}^{n} \left( y(x_i) - \hat{y}(x_i) \right)^2$$ (8)
When the response is binary, there are three criteria for the performance of CV: the misclassification rate, the Brier score and the Mean value of Minus Log-likelihood (MML), corresponding to the following formulae:

\[
CER_{\text{CV}} = \hat{\pi} \left[ I(\hat{\pi}_i < 1/2) + I(\hat{\pi}_i = 1/2) \right] + (1 - \hat{\pi}) \left[ I(\hat{\pi}_0 > 1/2) + I(\hat{\pi}_0 = 1/2) \right]
\]

\[
BRIER_{\text{CV}} = \sum_{i} (y_i - \hat{p}_{-i})^2
\]

\[
MML_{\text{CV}} = -\sum_{i} \left( y_i \log \hat{p}_{-i} + (1 - y_i) \log (1 - \hat{p}_{-i}) \right)
\]

where \(y_i, i = 1, \ldots, n\), denotes the \(i\)-th binary response and \(\hat{p}_{-i}\) is the prediction of \(y_i\) when the \(i\)-th sample unit is removed from the data and logistic distribution is used.

The optimum regularization parameter is the one that minimizes the error rate measure chosen as the CV criterion.

It is shown that the rescaled optimum error rates for all three types considered here are more or less coincident (Van Houwelingen and le Cessie (1990), p.1308).

In any way, LOO CV often works well for estimating generalization error for continuous error functions such as the mean squared error, but it may perform poorly for discontinuous error functions such as the number of misclassified cases. In the latter case k-fold is preferred, but if k gets too small, the error estimate is pessimistically biased because of the difference in training-set size between the full-sample analysis and the cross validation analysis (Shao, 1993).

5. STATISTICAL ANALYSIS OF “OC” DATA-SET

The data-set was obtained from a case-control study carried out at the Sector of Oral Medicine (University of Palermo, Italy). The response variable is the onset of OC, defined as dichotomous (presence/absence). The explicative variables are the socio demographic variables of age, gender and smoking, the local variables are Denture (D) and Hyposalivation/Xerostomia (IPX) and the systemic variable is Immunodeficiency non–HIV related (IM-).

The study group was composed of 89 patients affected by OC, clinically and microbiologically diagnosed (Campisi, Pizzo et al. 2002). The control group was composed of 98 individuals without any clinical sign of OC and matched for age and gender with the study group. After recruitment in a secondary epidemiological hospital base, where the output (onset of OC) was taken into account, it was found that in the control group local and systemic variables were absent. Briefly, medical
clinical research is very often carried out in a secondary base, where the cases are defined before the base is identified. In this approach, e.g., in a hospital-based study, the base is identified as the source of cases, and the cases are all patients diagnosed with the study disease; controls should also derive from the base, as individuals who would have become study cases if they had developed the disease in question at the time of investigation. In accordance with these statements, in our study all the cases were subjects experiencing OC, and the controls were used to estimate risk distribution in the catchment.

Firstly, we used the SAS/STAT empirical approach to detect data separation. Secondly, we fitted the MEL estimator with $\delta$ chosen on the base of four different methods: IRT, HOS-CV, 10-fold CV and LOO-CV. As criteria for evaluating the performance of CV we used $\text{BRIER}_{\text{CV}}$ and $\text{MML}_{\text{CV}}$, which are the most used in literature on the subject.

5.1 DETECTING SEPARATION

In Tab. 1 it can be seen that quasi-complete separation in data arises in correspondence with the variables of Denture, IPX and IM-.

The last change in the log-likelihood function in SAS is 0.0001356523 and the asymptotic dispersion matrix is unbounded with the diagonal element for Denture, IM- and IPX respectively, equal to 18485.81, 32950.45, 13615.14. Within the first eight iterations, the parameters for age, Sex and Smoking stabilize. The same does not happen for Denture, IPX and IM-.

<table>
<thead>
<tr>
<th>Iter</th>
<th>-2LogL</th>
<th>Intercept</th>
<th>SEX</th>
<th>AGE</th>
<th>SMOKE</th>
<th>DENTURE</th>
<th>IPX</th>
<th>IM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>258.804</td>
<td>0.096</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>116.202</td>
<td>-1.679</td>
<td>0.010</td>
<td>0.477</td>
<td>0.230</td>
<td>0.754</td>
<td>0.963</td>
<td>0.945</td>
</tr>
<tr>
<td>2</td>
<td>85.655</td>
<td>-3.428</td>
<td>0.049</td>
<td>0.623</td>
<td>0.350</td>
<td>1.403</td>
<td>1.667</td>
<td>1.781</td>
</tr>
<tr>
<td>3</td>
<td>75.052</td>
<td>-5.237</td>
<td>0.087</td>
<td>0.727</td>
<td>0.457</td>
<td>2.040</td>
<td>2.326</td>
<td>2.504</td>
</tr>
<tr>
<td>4</td>
<td>71.738</td>
<td>-6.871</td>
<td>0.099</td>
<td>0.800</td>
<td>0.529</td>
<td>2.570</td>
<td>2.872</td>
<td>3.087</td>
</tr>
<tr>
<td>5</td>
<td>70.601</td>
<td>-8.414</td>
<td>0.097</td>
<td>0.821</td>
<td>0.548</td>
<td>3.075</td>
<td>3.383</td>
<td>3.611</td>
</tr>
<tr>
<td>6</td>
<td>70.192</td>
<td>-9.926</td>
<td>0.097</td>
<td>0.823</td>
<td>0.550</td>
<td>3.578</td>
<td>3.886</td>
<td>4.117</td>
</tr>
<tr>
<td>7</td>
<td>70.043</td>
<td>-11.431</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>4.079</td>
<td>4.387</td>
<td>4.618</td>
</tr>
<tr>
<td>8</td>
<td>69.988</td>
<td>-12.932</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>4.579</td>
<td>4.888</td>
<td>5.119</td>
</tr>
<tr>
<td>9</td>
<td>69.968</td>
<td>-14.433</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>5.080</td>
<td>5.388</td>
<td>5.619</td>
</tr>
<tr>
<td>10</td>
<td>69.961</td>
<td>-15.933</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>5.580</td>
<td>5.888</td>
<td>6.119</td>
</tr>
<tr>
<td>11</td>
<td>69.958</td>
<td>-17.433</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>6.080</td>
<td>6.388</td>
<td>6.619</td>
</tr>
<tr>
<td>12</td>
<td>69.957</td>
<td>-18.933</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>6.580</td>
<td>6.888</td>
<td>7.119</td>
</tr>
<tr>
<td>13</td>
<td>69.956</td>
<td>-20.433</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>7.080</td>
<td>7.388</td>
<td>7.619</td>
</tr>
<tr>
<td>14</td>
<td>69.956</td>
<td>-21.933</td>
<td>0.096</td>
<td>0.823</td>
<td>0.550</td>
<td>7.580</td>
<td>7.888</td>
<td>8.119</td>
</tr>
</tbody>
</table>
5.2 MEL ESTIMATION AND METHODS OF CHOICE FOR $\delta$

If we perform the MEL estimator choosing $\delta$ using the IRT method, it becomes evident that the value which stabilizes the curve is equal to 0.055. For sake of evidence, we built the ridge trace on the base of the absolute variation $\Delta(\delta) = \hat{\beta}(\delta + i) - \hat{\beta}(\delta)$, with $i = 0.005$ and $\delta = \varepsilon > 0$ respectively, for D, IM- and IPX.

The estimated regression coefficient and the 95% Confidence intervals obtained on the base of 1000 bootstrap iterations are shown in Table 2.

### Tab. 2. MEL Estimates of regression coefficients and Odds Ratios.

<table>
<thead>
<tr>
<th></th>
<th>Bootstrap Estimate</th>
<th>Bias</th>
<th>SE</th>
<th>95% CI lower limit</th>
<th>95% CI upper limit</th>
<th>OR</th>
<th>95% CI OR lower limit</th>
<th>95% CI OR upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>-2.65</td>
<td>-0.07</td>
<td>0.48</td>
<td>-3.52</td>
<td>-1.64</td>
<td>0.07</td>
<td>0.03</td>
<td>0.19</td>
</tr>
<tr>
<td>SEX</td>
<td>0.14</td>
<td>-0.01</td>
<td>0.45</td>
<td>-0.73</td>
<td>1.03</td>
<td>1.15</td>
<td>0.48</td>
<td>2.80</td>
</tr>
<tr>
<td>AGE</td>
<td>1.30</td>
<td>0.02</td>
<td>0.57</td>
<td>0.10</td>
<td>2.28</td>
<td>3.67</td>
<td>1.11</td>
<td>9.78</td>
</tr>
<tr>
<td>SMOKE</td>
<td>0.78</td>
<td>-0.01</td>
<td>0.55</td>
<td>-0.41</td>
<td>1.81</td>
<td>2.18</td>
<td>0.66</td>
<td>6.11</td>
</tr>
<tr>
<td>DENTURE</td>
<td>3.83</td>
<td>0.06</td>
<td>0.49</td>
<td>2.79</td>
<td>4.74</td>
<td>46.06</td>
<td>16.28</td>
<td>114.43</td>
</tr>
<tr>
<td>IM-</td>
<td>4.70</td>
<td>0.08</td>
<td>0.45</td>
<td>3.22</td>
<td>5.29</td>
<td>109.95</td>
<td>25.03</td>
<td>198.34</td>
</tr>
<tr>
<td>IPX</td>
<td>4.35</td>
<td>0.09</td>
<td>0.40</td>
<td>3.43</td>
<td>4.96</td>
<td>77.48</td>
<td>30.88</td>
<td>142.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>-3.07</td>
<td>-0.18</td>
<td>0.70</td>
<td>-4.14</td>
<td>-1.73</td>
<td>0.05</td>
<td>0.02</td>
<td>0.18</td>
</tr>
<tr>
<td>SEX</td>
<td>0.20</td>
<td>0.02</td>
<td>0.61</td>
<td>-1.00</td>
<td>1.52</td>
<td>1.22</td>
<td>0.37</td>
<td>4.57</td>
</tr>
<tr>
<td>AGE</td>
<td>0.98</td>
<td>0.03</td>
<td>0.76</td>
<td>-0.88</td>
<td>2.37</td>
<td>2.66</td>
<td>0.41</td>
<td>10.70</td>
</tr>
<tr>
<td>SMOKE</td>
<td>1.51</td>
<td>0.04</td>
<td>0.75</td>
<td>-0.07</td>
<td>2.93</td>
<td>4.53</td>
<td>0.93</td>
<td>18.73</td>
</tr>
<tr>
<td>DENTURE</td>
<td>5.06</td>
<td>0.19</td>
<td>0.59</td>
<td>3.65</td>
<td>5.85</td>
<td>157.59</td>
<td>38.47</td>
<td>347.23</td>
</tr>
<tr>
<td>IM-</td>
<td>5.61</td>
<td>0.21</td>
<td>0.52</td>
<td>4.40</td>
<td>6.25</td>
<td>273.14</td>
<td>81.45</td>
<td>518.01</td>
</tr>
<tr>
<td>IPX</td>
<td>6.04</td>
<td>0.20</td>
<td>0.57</td>
<td>4.42</td>
<td>6.65</td>
<td>419.89</td>
<td>83.10</td>
<td>772.78</td>
</tr>
</tbody>
</table>
The application of the three CV methods, HOS-CV, 10-fold CV and LOO-CV, using the Brier score and the MML$_{CV}$ as CV criteria, give similar results: the value for $\delta$ that minimizes both criteria for 10-fold-CV is $\delta = 0.015$, while for LOO-CV $\delta$ results a little larger: based on the brier score, $\delta = 0.025$ while based on the MML$_{CV}$ criterion, $\delta = 0.025$. Finally, HSO-CV indicates $\delta = 0.02$, based on the brier score and $\delta = 0.015$, based on MML$_{CV}$.

Therefore, all the Cross Validation criteria agree, suggesting that $\delta = 0.02$, seemingly rather less than the value suggested by the IRT method. The estimated regression coefficient and the 95% Confidence intervals obtained on the base of 1000 bootstrap iterations also are in Table 2.

6. DISCUSSION

The OC data-set resulted quasi-completely separated with respect to Denture, IM-and IPX variables.

Fig. 3: MML$_{CV}$ and Brier score for 10-fold, HSO and LOO.
When substituting observed data with pseudo-data corrected for misclassification, the parameter estimates, obtained using MEL method, become finite. With both IRT and CV methods, MEL odds ratio estimates turn out to be very large and imprecise: for D, IM- and IPX variables, the IRT method yields OR=46.06, OR=109.95 and OR=77.48 respectively. For the same variables, the CV method gives higher results (OR=157.59, OR=273.44, OR=419.89). Moreover, such estimates have large standard errors (0.49, 0.45 and 0.40 by IRT and 0.59, 0.52 and 0.57 by CV for D, IM- and IPX respectively). Even the bias, computed as the difference between the observed MEL estimates and the mean of the one thousand MEL bootstrap replicates, is generally higher for the $\delta$ obtained with CV (0.19 vs 0.06 for D, 0.21 vs 0.08 for IM-, 0.0 vs 0.09 for IPX). However, such a result cannot be interpreted as the IRT method being more precise than the CV because the bias is estimated as a deviation from a sample estimate.

In brief, estimates obtained using the IRT and CV methods both suffer from undesirable properties which naturally ensue as a consequence of data separation. Moreover, the more the classification rate tends to zero the more the MEL estimates tend to infinity.

Such defects are unavoidable unless we compute the MEL estimator on a pseudo-configuration that is extremely distant from the observed separated configuration.

It is possible to recognize two main advantages of using CV: the first is that 10-fold, HSO and LOO all agree, implying a misclassification rate which is much less than the one suggested by IRT. In other words, CV approximates the observed data-set better, yielding a pseudo-configuration which is more consistent with the observed amount of separation existent within the data. The major width of confidence intervals obtained using CV, compared to IRT, is a consequence of the greater parameter size, which rises by the smaller error rate.

The second advantage is that CV estimates the regularization parameter using a multivariate process that takes account of all the variables together, in order to predict the observed values. On the contrary, IRT considers the ridge trace of one variable at a time.

When choosing between IRT and CV, it is necessary to look for a good compromise between the need to obtain finite odds ratios and the need to obtain the maximum proximity between the observed and the pseudo data configuration.

Results from OC data-set analysis show that CV appears to answer this requirement better than IRT. From the clinical point of view, the MEL estimator allows the identification of variables that otherwise are not investigable statistically, so that they can be taken into account in medical practice. Only Cross Validation
allows a better approximation of the reality, as shown in daily clinical experience. Indeed, the different variables investigated are often cited in a fragmented way, but never, until now, in our knowledge, by investigating the statistical relationships which regulate them.

CONCLUDING REMARKS

In this paper, we have presented a method for dealing with quasi-completely separated data-sets based on the hidden logistic framework. The aim was to modify the maximum likelihood method to correct misclassification in data. More specifically, we adopted the MEL estimator of Rousseeuw and Christmann (2003) but at the same time introduced a Cross-validation based criterion for the choice of regularization parameter, in order to guarantee the uniqueness of the solution.

From the Cross-Validation criteria, we considered 10-fold, LOO-CV and HSO-CV methods and evaluated their performance by comparing them with an inspection of the ridge trace. All the CV criteria were in agreement, each suggesting one value for the regularization parameter that was much lower than the one suggested using the graphical method. Even though MEL estimates based on Cross-Validation have larger standard errors than those obtained through the inspection of the ridge trace, this is the obvious effect of considering a lower margin of approximation when substituting observations with pseudo-observations. Therefore, if the main aim of the analysis is to obtain maximum likelihood estimates with minimum misclassification error (as long as these are consistent with the observed amount of separation existing in the data), we obtain better results using the Cross-Validation method than using graphical inspection of the ridge trace.

Current analysis shows that the MEL estimator allows the clinical identification of important variables, which are not otherwise investigable statistically. Further research on separated data configuration must be carried out in order to reduce the width of CI for regression coefficients and corresponding OR estimates.

REFERENCES


LA STIMA DELL’ODDS RATIO IN PRESENZA DI DATI COMPLETAMENTE O QUASI COMPLETAMENTE SEPARATI

Riassunto

In presenza di dati completamente o quasi completamente separati non esistono stime di massima verosimiglianza del modello di regressione logistica. In ambito medico, la questione è di notevole rilievo per la necessità di ottenere stime finite dell’odds ratio. I software statistici non risolvono il problema della stima con dataset non sovrapposti. In questo lavoro, proponiamo di applicare il modello di regressione logistica “nascosta” e lo stimatore MEL di Rousseeuw and Christmann (2003, per il quale si ottiene un’unica soluzione tramite ispezione grafica della traccia ridge dei parametri di regressione (IRT). In alternativa, suggeriamo di scegliere il parametro di regolarizzazione mediante un metodo basato sulla Cross validation (CV). I metodi sono validati su un dataset relativo a uno studio caso-controllo sulla candidosi orale. In base alla nostra analisi, il metodo CV determina un errore di misclassificazione inferiore rispetto al metodo IRT.