ROBUST ANALYSIS OF CLUSTERED BINARY DATA: BETA-BINOMIAL MODEL UNDER RESPONSE MISCLASSIFICATION

Maxim Pashkevich, Yuriy Kharin

Department of Mathematical Modeling and Data Analysis, Faculty of Applied Mathematics and Computer Science, Belarussian State University, 4 F. Skaryny Ave., Minsk, Belarus, e-mails: pma@omegasoftware.com, kharin@bsu.by

Abstract. New estimators and predictors for the beta-binomial model that are robust to binary data misclassifications are proposed. It is proved that the proposed estimators are consistent and the proposed predictors have the minimum mean square error in the case of known distortion levels. The efficiency of the methods is illustrated on real data from mediaplanning.

Introduction

Methods for clustered binary data that arises in longitudinal studies assume that binary responses are measured without errors. However, recent research shows that in practice this may not be the case [8]. Misclassified binary responses arise in mediaplanning in TV ratings measurements [4], in medicine when the presence or absence of a medical condition is identified through an imperfect diagnostic test [8], in sociology when modeling panel behavior [5], in econometric modeling [2] and in other fields [8]. It makes the problem of developing robust to binary misclassifications statistical methods for analyzing binary clustered data important for practical applications.

Beta-binomial model (BBM) is often used to model clustered binary data in the case when there is no additional information on observations conditions available. To estimate the parameters of the BBM, the method of moments (MM) and the method of maximum likelihood (MML) are usually used [9]. The forecasting based on the BBM is performed using the empirical Bayes predictor that has the minimum mean square error when there are no distortions [1, 7]. However, under distortions the described above estimation and prediction methods loose their performance [3, 6].

In this paper, new estimators and predictors for the BBM that are robust to binary data misclassifications are proposed. It is proved that the new estimators are consistent and the new predictors have the minimum mean square error in the case of known distortion levels. The efficiency of the proposed estimation and forecasting methods is illustrated via test on real data from mediaplanning.

Forecasting problem for binary distorted data

Let us have $K$ objects and a random event $A$ and let $B$ be a binary $(K \times n)$-matrix that contains the results of $n$ Bernoulli experiments with the event $A$ over the objects: $B = (b_{ij})$, $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots, n$, $b_{ij} = 1(0)$. Let us make the following two assumptions.

$A_1$. The probability properties of the objects are stable in time.

$A_2$. For the object $i$, the random probability of success $p_i$ has the beta distribution with true unknown parameters $\alpha_0$, $\beta_0$; random variables $p_1$, $p_2$, $\ldots$, $p_K$ are independent in total.

Suppose that the binary matrix $B$ is distorted by random binary errors $\{\eta_{ij}\}$:
$$\tilde{b}_y = b_y \oplus \eta_y, \quad P\{\eta_y = I|b_y = I\} = \varepsilon_i, \quad \varepsilon_i \in [0, \varepsilon_i, \varepsilon_i, 1/2], \quad l = 0, 1, \quad (1)$$

where $\oplus$ is the exclusive logical XOR operator, and $\{\eta_y\}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, K$, are independent Bernoulli variables. Let us refer to $\varepsilon_0, \varepsilon_1$ as distortion levels.

The problem is to forecast the unknown success probabilities $p_1, p_2, \ldots, p_K$ having the distorted sample $X$ that is calculated using the distorted matrix $\tilde{B}$ as

$$x_i = \sum_{j=1}^{n} \tilde{b}_{ij}, \quad i = 1, 2, K, K. \quad (2)$$

**Robust estimation of the beta-binomial model parameters**

It can be proved that the sample $X$ is a random sample from the discrete distribution

$$p_r(n, \alpha, \beta, \varepsilon_0, \varepsilon_1) = P\{x = r\} = \sum_{i=0}^{n} w_r(n, \varepsilon_0, \varepsilon_1) \cdot p_0^n(n, \alpha, \beta), \quad r = 0, 1, K, n, \quad (3)$$

where $W(w_n)$ is a $(n+1) \times (n+1)$ matrix that is defined as

$$w_r(n, \varepsilon_0, \varepsilon_1) = \sum_{r=\max(1, r)}^{\min(n, r)} C_r^{n-\varepsilon_0} C_{n-r}^{\varepsilon_0} (1-\varepsilon_1)^{r-1} \varepsilon_1^{n-r} (1-\varepsilon_1)^{n-r}, \quad r = 0, 1, K, n, \quad (4)$$

and $p^0_r(n, \alpha, \beta)$ are the classical BBM probabilities [9]. Let us refer to the model that corresponds to this distribution as to the distorted beta-binomial distribution (DBBM).

The Inconsistency of the MM and MML Estimators Under Distortions. The results from [6] allow us to conclude that in the case of distorted sample, the classical MM-estimators and MML-estimators become inconsistent. In the following subsections, new robust estimators of the BBM parameters are introduced. Two cases are considered: the case of known distortion levels $\varepsilon_0, \varepsilon_1$ and the general case when there is no priori information.

**Robust Estimation in the Case of Known Distortion Levels.** In practice, it is a common situation when the estimates of distortion levels are known in advance. The source of prior information can be the past work experience or experts knowledge. In this case, it is possible to develop new methods of the beta-binomial model parameters estimation that lead to consistent estimators of $\alpha, \beta$ parameters.

**Theorem 1.** The consistent MM-estimators of the BBM parameters $\alpha, \beta$ under distortions (1) with known levels $\varepsilon_0, \varepsilon_1$ are defined as

$$\hat{\alpha}_{MM}(\varepsilon_0, \varepsilon_1) = \frac{(m^*_1 - \varepsilon_0 n)(m^*_1 n - m^*_0 + \varepsilon_0(n-1))(m^*_1 - (1 - \varepsilon_1)n) - m^*_1 \varepsilon_1(n-1))}{(1 - \varepsilon_0 - \varepsilon_1)(m^*_1 n - m^*_0 n - m^*_1 n^2(n-1))}, \quad (5)$$

$$\hat{\beta}_{MM}(\varepsilon_0, \varepsilon_1) = \frac{(m^*_1 - (1 - \varepsilon_1)n)(m^*_2 + m^*_1 \varepsilon_1(n-1) - m^*_1 n - \varepsilon_0(n-1))(m^*_1 - n(1 - \varepsilon_1))}{(1 - \varepsilon_0 - \varepsilon_1)(m^*_2 n - m^*_1 n - m^*_1 n^2(n-1))}, \quad (6)$$

$$m^*_1 = K^{-1} \sum_{i=1}^{K} x_i, \quad m^*_2 = K^{-1} \sum_{i=1}^{K} x_i^2. \quad (7)$$

**Remark.** A general estimation method of the BBM parameters $\alpha, \beta$, in the case of known distortion levels, is also proposed. Let $\tilde{p}(\varepsilon_0, \varepsilon_1)$ be a vector of empirical BBM probabilities for the distorted sample $X$. It is proved that one can filter the vector $\tilde{p}(\varepsilon_0, \varepsilon_1)$ using matrix $W(\varepsilon_0, \varepsilon_1)$ to get the empirical undistorted BBM probabilities: $\tilde{p}_0 = W^{-1}(\varepsilon_0, \varepsilon_1) \cdot \tilde{p}(\varepsilon_0, \varepsilon_1)$.

The problem of robust estimation is reduced to the classical BBM estimation problem.
Simultaneous Estimation of $\alpha, \beta, \varepsilon_0, \varepsilon_1$. Let $m_1^*, m_2^*, m_3^*, m_4^*$ be the first four moments for the distorted sample $X$. With the result of the Theorem 1, the problem of joint estimation of $\alpha, \beta, \varepsilon_0, \varepsilon_1$ is reduced to the solution of the non-linear system

$$
m_3^* = m_3(\alpha(\varepsilon_0, \varepsilon_1), \beta(\varepsilon_0, \varepsilon_1), \varepsilon_0, \varepsilon_1), \quad m_4^* = m_4(\alpha(\varepsilon_0, \varepsilon_1), \beta(\varepsilon_0, \varepsilon_1), \varepsilon_0, \varepsilon_1),$$

under the assumption $m_1^* = \text{const}, \ m_2^* = \text{const}$. Let $J_0^\text{cond}$ be a Jacobi matrix of (8) under this assumption, then the method of Newton leads us to the following iterative procedure:

$$
\begin{pmatrix}
\varepsilon_0^{k+1} \\
\varepsilon_1^{k+1}
\end{pmatrix} = \begin{pmatrix}
\varepsilon_0^k \\
\varepsilon_1^k
\end{pmatrix} + \lambda \cdot (J_0^\text{cond})^{-1} \begin{pmatrix}
m_3^* - m_3(\alpha(\varepsilon_0^k, \varepsilon_1^k), \beta(\varepsilon_0^k, \varepsilon_1^k), \varepsilon_0^k, \varepsilon_1^k) \\
m_4^* - m_4(\alpha(\varepsilon_0^k, \varepsilon_1^k), \beta(\varepsilon_0^k, \varepsilon_1^k), \varepsilon_0^k, \varepsilon_1^k)
\end{pmatrix},
$$

where $\lambda \in (0, 1)$ is the algorithm parameter that ensures convergence of the procedure in the case of large distortion levels $\varepsilon_0, \varepsilon_1$. The elements of the conditional Jacobi matrix $J_0^\text{cond}$ can be calculated using the results from the previous work [6].

Another approach of joint estimation employs the method of maximum likelihood. In this case, the estimation problem is reduced to the constrained maximization problem

$$
l(\alpha, \beta, \varepsilon_0, \varepsilon_1) = \sum_{r=0}^n f_r \cdot \ln(p_r, (\alpha, \beta, \varepsilon_0, \varepsilon_1)) \rightarrow \max, \quad \alpha > 0, \quad \beta > 0, \quad 0 \leq \varepsilon_0, \varepsilon_1 \leq 1,$$

where $\{f_r\}$ are the frequencies for the distorted sample $X$. The problem (10) is solved numerically using the modification of the steepest descent method that is based on the results from [6].

Robust forecasting based on the beta-binomial model

It is proved [6] that in the case of binary distortions (1), the classical Bayes predictor looses its optimality. As a result, a new Bayes-based predictor that enjoys this property has to be developed. Let us have some statistical estimates $\hat{\alpha}, \hat{\beta}, \varepsilon_0, \varepsilon_1$ of the BBM parameters and the distortion levels and let us denote $y^{[i+]} = y_{i+} \mathbb{I}(y+z-1), \ y \in \mathbb{R}, \ z \in \mathbb{N}$.

**Theorem 2.** In the case of known distortion levels $\varepsilon_0, \varepsilon_1$, the optimal forecast is

$$
\hat{P}_\text{bayer}(s) = E[p|s, f\hat{\alpha}, f\hat{\beta}] = \sum_{i=0}^n f_{i+} f_{\hat{\alpha} + \hat{\beta} + n},
$$

which provides the minimum mean square error

$$
r^2_\text{min}(\hat{P}_\text{bayer}) = \frac{(\alpha + j + \hat{\alpha} + i)}{(\alpha + \beta + n)^2} \sum_{i=0}^n f_{i+} f_{\hat{\alpha} + \hat{\beta} + n} \frac{C_n^i w_i B(\hat{\alpha} + i, \hat{\beta} + n - i)}{B(\hat{\alpha} + i, \hat{\beta} + n - i)},
$$

and has the following probability density function:

$$
f_p(x|s, f\hat{\alpha}, f\hat{\beta}) = \sum_{i=0}^n f_{i+} f_{i} (x), \quad L(f) = B(\hat{\alpha} + i, \hat{\beta} + n - i),
$$

where $s$ is the observed number of responses of the object, $p$ is its response probability and $\omega_i = C_n^i w_i B(\hat{\alpha} + i, \hat{\beta} + n - i)$.
Application of the developed methods to mediaplanning

The developed methods of robust estimation and prediction for the BBM have been implemented in the commercial software package Pinergy (Omega Software GmbH) for predicting the behavior of TV audience. To test the efficiency of the developed algorithms, real-life data for one of the German TV stations for the period 02.01.1999–01.01.2000 was used. Using this data, the developed DBBM was compared to the classical BBM for forecasting the Net Reach and Gross Rating Points (GRP) for master target groups of TV audience.

There were selected 11 TV commercial breaks that correspond to the fixed program (world news), weekday (Saturday) and day time (prime time). Tables 1–4 compare the adequacy of the classical BBD and the proposed DBBD for different estimation methods using Pearson’s $\chi^2$-criterion. As follows from the tables, the proposed DBBD together with the developed estimation algorithms significantly increases the modeling accuracy.

### Table 1. Adequacy of BBM (classical MM)

<table>
<thead>
<tr>
<th>Target group</th>
<th>M 14–29</th>
<th>M 30–49</th>
<th>M 50+</th>
<th>W 14–29</th>
<th>W 30–49</th>
<th>W 50+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value</td>
<td>0.41</td>
<td>0.96</td>
<td>0.01</td>
<td>0.82</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>Parameter $\alpha$</td>
<td>0.17</td>
<td>0.16</td>
<td>0.17</td>
<td>0.18</td>
<td>0.21</td>
<td>0.16</td>
</tr>
<tr>
<td>Parameter $\beta$</td>
<td>13.87</td>
<td>3.14</td>
<td>4.56</td>
<td>8.33</td>
<td>5.27</td>
<td>3.25</td>
</tr>
</tbody>
</table>

### Table 2. Adequacy of BBM (classical MML)

<table>
<thead>
<tr>
<th>Target group</th>
<th>M 14–29</th>
<th>M 30–49</th>
<th>M 50+</th>
<th>W 14–29</th>
<th>W 30–49</th>
<th>W 50+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value</td>
<td>0.45</td>
<td>0.97</td>
<td>0.02</td>
<td>0.80</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>Parameter $\alpha$</td>
<td>0.17</td>
<td>0.17</td>
<td>0.19</td>
<td>0.19</td>
<td>0.24</td>
<td>0.17</td>
</tr>
<tr>
<td>Parameter $\beta$</td>
<td>13.53</td>
<td>6.55</td>
<td>5.08</td>
<td>8.81</td>
<td>5.78</td>
<td>3.47</td>
</tr>
</tbody>
</table>

### Table 3. Adequacy of DBBM (the developed method of joint estimation of the parameters and the distortion levels based on MM)

<table>
<thead>
<tr>
<th>Target group</th>
<th>M 14–29</th>
<th>M 30–49</th>
<th>M 50+</th>
<th>W 14–29</th>
<th>W 30–49</th>
<th>W 50+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value</td>
<td>0.28</td>
<td>0.99</td>
<td>0.32</td>
<td>0.89</td>
<td>0.84</td>
<td>0.14</td>
</tr>
<tr>
<td>$\chi^2$-statistics</td>
<td>10.8993</td>
<td>0.9956</td>
<td>10.4146</td>
<td>4.3460</td>
<td>4.9850</td>
<td>13.4565</td>
</tr>
<tr>
<td>Parameter $\alpha$</td>
<td>0.09</td>
<td>0.13</td>
<td>0.14</td>
<td>0.13</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Parameter $\beta$</td>
<td>9.57</td>
<td>5.58</td>
<td>3.88</td>
<td>7.07</td>
<td>4.00</td>
<td>3.14</td>
</tr>
<tr>
<td>Distortion level $\epsilon_0$</td>
<td>0.003</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.006</td>
<td>0.001</td>
</tr>
<tr>
<td>Distortion level $\epsilon_1$</td>
<td>0.060</td>
<td>0.000</td>
<td>0.042</td>
<td>0.000</td>
<td>0.06</td>
<td>0.004</td>
</tr>
</tbody>
</table>

### Table 4. Adequacy of DBBM (the developed method of joint estimation of the parameters and the distortion levels based on MML)

<table>
<thead>
<tr>
<th>Target group</th>
<th>M 14–29</th>
<th>M 30–49</th>
<th>M 50+</th>
<th>W 14–29</th>
<th>W 30–49</th>
<th>W 50+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value</td>
<td>0.63</td>
<td>0.98</td>
<td>0.77</td>
<td>0.88</td>
<td>0.83</td>
<td>0.50</td>
</tr>
<tr>
<td>$\chi^2$-statistics</td>
<td>7.1145</td>
<td>2.4552</td>
<td>5.6728</td>
<td>4.3983</td>
<td>5.0365</td>
<td>8.3400</td>
</tr>
<tr>
<td>Parameter $\alpha$</td>
<td>0.11</td>
<td>0.15</td>
<td>0.10</td>
<td>0.15</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>Parameter $\beta$</td>
<td>9.12</td>
<td>6.39</td>
<td>3.14</td>
<td>7.51</td>
<td>4.31</td>
<td>2.36</td>
</tr>
<tr>
<td>Distortion level $\epsilon_0$</td>
<td>0.001</td>
<td>0.002</td>
<td>0.007</td>
<td>0.002</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>Distortion level $\epsilon_1$</td>
<td>0.048</td>
<td>0.015</td>
<td>0.068</td>
<td>0.008</td>
<td>0.020</td>
<td>0.111</td>
</tr>
</tbody>
</table>
Fig. 1. Reach/GRP curves: real data (solid line), the classical beta-binomial model (dashed line), the proposed distorted beta-binomial model (dotted line)

In practice, media-planners evaluate the adequacy of the customer’s respond model using Reach/GRP curves [4]. Fig 1 compares Reach/GRP curves for the classical BBM and the proposed DBBM with the real data curve. As follows from the figure, the proposed DBBM together with the developed estimation and forecasting algorithms ensures much more accurate approximation of the Reach/GRP function than the classical BBM.

References