A FIXED-INVERSE BINARY MISCLASSIFICATION MODEL UNDER POSSIBLE FALSE-POSITIVE MISCLASSIFICATION

Asmerom Tesfamichael, Advisor: Dr. Kent Riggs
Stephen F Austin State University – Department of Mathematics and Statistics

Abstract

• In this project, we develop a particular statistical model for binary data that allows for the possibility of false-positive misclassification. To account for the misclassification, the model incorporates a two-stage sampling scheme.

• Next, we apply maximum likelihood methods to find estimators of the primary prevalence parameter \( p \) as well as the false-positive misclassification rate parameter \( \phi \). In addition, we derive confidence intervals for \( p \) based on inverting Wald, score and likelihood ratio statistics.

• Also, we graphically compare coverage and width properties of the Wald-based, score-based, and likelihood ratio-based confidence intervals for \( p \) through a Monte Carlo simulation. The simulation study is done under different parameters and sample size configurations. Also, we apply the newly-derived confidence intervals for \( p \) to a real data set.

Introduction

Due to practical reasons such as cost and time savings, fallible classifiers which are prone to error are used to classify binary data.

• Misclassification may result in false-negatives or false-positives.

• Misclassification errors may distort results of statistical analysis.

• Models that account for misclassification have been developed to compensate for the effect of errors.

• The misclassification rate parameter is a measurable feature of a statistical model that accounts for misclassification.

• Different applications require different statistical models, which have specific advantages and limitations.

• Better estimation can be done by an infallible device, but at a higher cost.

• A double sampling scheme using both fallible and infallible devices may be used at a reasonable cost, while properly accounting for misclassification.

• We consider a model that allows only for false-positive misclassification, which treated the first sample of a two-stage sampling scheme as fixed and the second stage of the scheme as random (inverse sampling).

Two-Stage Sampling Scheme

The double sampling scheme involves the use of a fallible classier (cheap) and infallible classifyer in two stages in an effort to appropriately estimate \( p \) and the matrix of misclassification. The first stage involves the use of a fallible classifier that is prone to producing false-positives under fixed sampling. The second stage involves using an infallible fallible classifier under inverse sampling. For insight into this two-stage scheme consider the following example (* denotes false-positive):

<table>
<thead>
<tr>
<th>Stage</th>
<th>Pop.</th>
<th>Fixed</th>
<th>Fallible</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Fixed-Inverse Binary Misclassification Model

For the two-stage sampling scheme define the following quantities:

- \( y = \# \) of observations labeled success after \( n_0 \) trials of the fallible device in stage 1.
- \( n_0 = \# \) of observations labeled failure by both fallible and infallible devices in stage 1.
- \( n_2 = \# \) of observations labeled success by the fallible device but not by the infallible device in stage 2.
- \( n_2 = \# \) of observations labeled success by both fallible and infallible devices in stage 2.

For example above, \( y = 5, n_0 = 1, n_1 = 2, \) and \( n_2 = 5 \).

Distributional Assumptions

The Binomial distribution and Negative Multinomial distribution are used to model the counts \( y, n_0, n_2 \):

\[
f(0) = \binom{n}{y} p^y (1-p)^{n-y}
\]

where

\[
f(n_0, n_2) = \frac{n_0! n_2!}{(n_0 + n_2)!} (1-p)^{n_0} (1-\phi)^{n_2} \phi^{n_1}
\]

and

\[
f(n_0, n_1, n_2) = \frac{n_0! n_1! n_2!}{(n_0 + n_1 + n_2)!} (1-p)^{n_1} (1-\phi)^{n_2} \phi^{n_1}
\]

Maximum Likelihood Estimators

\[
\hat{p} = \frac{n_1}{n_0 + n_1 + n_2} \text{ and } \hat{\phi} = \frac{n_1}{n_0 + n_1 + n_2}
\]

Likelihood CI values of \( p \) that satisfy:

\[
2 \left( f(p; \phi) - f(p; \phi_0) \right) \leq \chi^2(1)(\alpha)
\]

where

\[
Z_0 = (1 - \alpha/2) \text{ percentile for the standard normal distribution}
\]

\[
v_f(p; \phi) = \frac{\partial f(p; \phi)}{\partial p}
\]

\[
y_f(p; \phi) = (1 - \alpha) \text{ percentile of a chi-squared distribution with one degree of freedom}
\]

\[
\Phi(p; \phi) = (1, 1) \text{ element of the inverse of Fisher’s information matrix}
\]

Motivating Example

• The western blot procedure (WBP) is one diagnostic test of the herpes simplex virus. Out of 693 women tested, the WBP yielded that 373 had the virus. (Hildeshiem and Boese)

• Under the binomial model (not accounting for misclassification), the maximum likelihood estimator (MLE) and Wald confidence interval for \( p \) is

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Wald CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>0.541</td>
<td>(0.504, 0.578)</td>
</tr>
</tbody>
</table>

Simulation Results

• The estimate of \( p \) under Fixed-Inverse Binary Misclassification Model is smaller than under the binomial model, which is intuitive because the misclassification rate (\( \phi \)) is around 12%. Hence, the estimate of \( p \) under the binomial model is likely overestimated due to false-positives generated by using only the fallible classifier.

Simulation Results When \( n_2 = 0.05m \)

Estimated Actual Coverages and Actual Widths when \( n_1 = 0.05m \) and \( p = 0.25, \phi = 0.05 \)

Simulation Results When \( n_2 = 0.4m \)

Estimated Actual Coverages and Actual Widths when \( n_1 = 0.4m \) and \( p = 0.25, \phi = 0.05 \)

Bibliography


Printed by MegaPrint Inc., www.megaprint.com