Misclassification in Logistic Regression with Discrete Covariates

Ori Davidov*, David Faraggi and Benjamin Reiser
Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905 Israel

Summary

We study the effect of misclassification of a binary covariate on the parameters of a logistic regression model. In particular we consider $2 \times 2 \times 2$ tables. We assume that a binary covariate is subject to misclassification that may depend on the observed outcome. This type of misclassification is known as (outcome dependent) differential misclassification. We examine the resulting asymptotic bias on the parameters of the model and derive formulas for the biases and their approximations as a function of the odds and misclassification probabilities. Conditions for unbiased estimation are also discussed. The implications are illustrated numerically using a case control study. For completeness we briefly examine the effect of covariate dependent misclassification of exposures and of outcomes.

Key words: Asymptotic bias; Binary data; Differential misclassification; Logistic regression.

1. Introduction

It is well known that misclassification of either outcomes or covariates has important implications on parameter estimates and statistical inference. Bross (1954) was the first to investigate the effect of misclassification on binomial probabilities. Later research extended his investigation to misclassified $2 \times 2$ tables (Newell, 1963; Koch 1969; Goldberg 1975). The validity and power of tests for independence in the presence of misclassification has been studied by various authors (Mote and Anderson, 1965; Assakul and Proctor, 1967; Chiacchierini and Arnold, 1977). They conclude that the size of the test is unchanged but its power is usually diminished. More generally, the effect of misclassification on the association between exposures and outcomes has been investigated by numerous epidemiologists (e.g., Copeland et al., 1977 and the references therein).

In this communication the effect of misclassification on the parameters of a logistic regression is investigated. More specifically we derive new, exact and approximate, formulas for the asymptotic bias (on the log-odds scale) as a function of the odds and misclassification probabilities. We focus on $2 \times 2 \times 2$ tables and highlight the effect of a misclassified exposure on the regression parameters asso-

* Corresponding author: davidov@stat.haifa.ac.il
associated with other covariates. Previous studies of this problem have been primarily numeric (Greenland, 1980; Marshall and Hastrup, 1995).

2. Notation

Let \( Y \) denote a binary outcome associated with two exposures, or covariates, which we denote \( E \) and \( Z \). At present we assume that the covariates are binary. Their joint probability function is

\[
\pi_{ijk} = \Pr [Y = i, E = j, Z = k]
\]

where \( i, j, k = 0, 1 \). Let \( \pi = (\pi_{ijk}) \) denote the full vector of probabilities. The conditional distribution of \( Y \) is denoted \( \pi_{i|jk} \). It may be parameterized as

\[
\Pr [Y = 1 | E, Z] = \frac{\exp (\beta_0 + \beta_1 E + \beta_2 Z + \beta_3 EZ)}{1 + \exp (\beta_0 + \beta_1 E + \beta_2 Z + \beta_3 EZ)}
\]

which is a saturated logistic regression model (Agresti, 1990). This model is widely used in biostatistical applications to analyze both prospective and retrospective designs. We assume that the true value of the exposure \( E \) is generally unobserved. Instead we observe the random variable \( X \) which is a misclassified version of \( E \). They are related by

\[
0_i = \Pr [X = 1 | E = 1, Y = i],
\]

\[
\phi_i = \Pr [X = 0 | E = 0, Y = i]
\]

where \( i = 0, 1 \). Let \( \theta = (0_0, 0_1) \) and define \( \phi \) similarly. Note that the misclassification probabilities are expressed in terms of sensitivity and specificity and depend on the value of the outcome \( Y \). This is known as (outcome dependent) differential misclassification. The distribution for the observed data is denoted \( p \), where

\[
p_{ijk} = \Pr [Y = i, X = j, Z = k].
\]

In practice the model

\[
\Pr [Y = 1 | X, Z] = \frac{\exp (\gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ)}{1 + \exp (\gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ)}
\]

and not (1) is fit to the observed data and the parameter \( \gamma \) is estimated. We note that both saturated logistic models (1) and (4) hold simultaneously in the presence of misclassification; the first describes an unobserved probability distribution while the second describes an observed probability distribution. Finally define

\[
\Delta_k(\theta, \phi) = \beta_k - \gamma_k
\]

for \( k = 0, 1, 2, 3 \) to be the asymptotic bias induced by misclassification. This quantity measures the performance of the estimates under misclassification. Large values of \( \Delta_k \) indicate that misclassification has a large effect on the estimation of \( \beta_k \).
3. Asymptotic Bias

In this section we compute the asymptotic bias, on the log-odds scale, due to misclassification of an exposure. First we relate the regression parameters associated with (1) and (4) to the probabilities \( p \) and \( p \). Varying the value of the covariates \( E \) and \( Z \) in (1) we obtain the equations

\[
\begin{align*}
\pi_{1|00} &= \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}, \\
\pi_{1|10} &= \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}, \\
\pi_{1|01} &= \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}, \\
\pi_{1|11} &= \frac{\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)}.
\end{align*}
\]

Solving for \( \beta \) using the identity

\[
\frac{\pi_{i|jk}}{1 - \pi_{i|jk}} = \frac{\pi_{ijk}}{\pi_{i'j'k}}
\]

where \( i' + i = 1 \), we obtain the well known relationships

\[
\begin{align*}
\beta_0 &= \log \left( \frac{\pi_{100}}{\pi_{000}} \right), \\
\beta_1 &= \log \left( \frac{\pi_{110}}{\pi_{010}} \frac{\pi_{000}}{\pi_{100}} \right), \\
\beta_2 &= \log \left( \frac{\pi_{101}}{\pi_{001}} \frac{\pi_{000}}{\pi_{100}} \right), \\
\beta_3 &= \log \left( \frac{\pi_{111}}{\pi_{011}} \frac{\pi_{010}}{\pi_{000}} \frac{\pi_{001}}{\pi_{110}} \frac{\pi_{101}}{\pi_{100}} \right).
\end{align*}
\]

Similar expressions are obtained for the regression parameters \( \gamma \) in terms of the probabilities \( p \). It is clear that the \( p' \)'s are related to the \( \pi' \)'s through the misclassification probabilities \( \theta \) and \( \phi \). A straightforward conditional probability argument applied to (2–3) implies that

\[
\begin{align*}
p_{i1k} &= \pi_{i1k}\theta_i + \pi_{i0k}(1 - \phi_i), \\
p_{i0k} &= \pi_{i0k}\phi_i + \pi_{i1k}(1 - \theta_i)
\end{align*}
\]

for \( i, k = 0, 1 \). Hence the parameters \( \gamma \) are expressible as a function of \( \pi, \theta \) and \( \phi \).
3.1 Exact and approximate formulas for the bias

We start with the bias associated with the intercept. Equation (5) provides an expression for \( b_0 \) in terms of the probabilities \( p \). A similar expression in terms of \( p \) is available for \( \gamma_0 \). Since \( p \) is a function of \( \pi, \theta \) and \( \phi \), as in (9–10), so is \( \gamma_0 \). More formally we may write,

\[
\Delta_0(\theta, \phi) = \beta_0 - \gamma_0 = \log (R_0)
\]

where

\[
R_0 = \frac{\phi_0}{\phi_1} \left( \frac{1 + \xi_{00}}{1 + \xi_{10}} \right) \frac{1 - \theta_0}{1 - \theta_1},
\]

and

\[
\xi_{ik} = \frac{\pi_{1ik}}{\pi_{0ik}} = \frac{\Pr [Y = i, E = 1, Z = k]}{\Pr [Y = i, E = 0, Z = k]} = \frac{\Pr [E = 1 | Y = i, Z = k]}{\Pr [E = 0 | Y = i, Z = k]}
\]

is the retrospective odds of exposure given \( Y = i \) and \( Z = k \).

Note that the bias is zero whenever \( R_0 = 1 \). Conditions for unbiasedness are an easy consequence. For example if \( \theta_0 = \theta_1 \), \( \phi_0 = \phi_1 \) and \( \xi_{00} = \xi_{10} \) then \( R_0 = 1 \). This scenario corresponds to non differential misclassification and an odds ratio of one (i.e., \( \beta_1 = 0 \)) when \( Z = 0 \). Another interesting situation where the bias is zero arises with non differential misclassification and unit sensitivity (i.e., \( \theta_0 = \theta_1 = 1 \) and \( \phi_0 = \phi_1 \)) regardless of the value of the odds. Assuming non-differential misclassification and expanding \( \Delta_0 \) in a Taylor series, we obtain

\[
\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \xi_{00}^k (1 - e^{k\beta_1}) \left( \frac{1 - \theta_0}{\phi} \right)^k.
\]

It is easy to see that the bias is positive when \( \beta_1 < 0 \), negative when \( \beta_1 > 0 \) and identically zero if \( \beta_1 = 0 \), i.e., whenever the outcome \( Y \) is independent of the exposure \( E \). The behavior of the bias as a function of the misclassification probabilities, is easily obtained from the first term in the Taylor expansion and equals

\[
(\phi_0 - \phi_1) + \xi_{00}(1 - \theta_0) - \xi_{10}(1 - \theta_1).
\]

Next we calculate \( \Delta_1 \) which is the bias in estimating \( \beta_1 \) when the covariate associated with it, \( E \), is misclassified. Using (6) and (9–10) we find that

\[
\Delta_1(\theta, \phi) = \beta_1 - \gamma_1 = \log (R_1)
\]

where

\[
R_1 = \frac{\theta_0 \phi_1}{\theta_1 \phi_0} \left[ \frac{\xi_{00}^{-1}(1 - \phi_0)/\theta_0, \xi_{10}(1 - \theta_1)/\phi_1]^{\otimes}}{\xi_{00}(1 - \theta_0)/\phi_0, \xi_{10}^{-1}(1 - \phi_1)/\theta_1]^{\otimes}} \right],
\]

(12)
and \([a_1, \ldots, a_m] \equiv \prod_i (1 + a_i)\). The general formula for the series expansion of \(\Delta_1\) is quite complicated; however its first order approximation has the form

\[
(\theta_0 - \theta_1) + (\phi_1 - \phi_0) + \xi_{10}(1 - \theta_1) - \xi_{00}(1 - \theta_0) \\
+ \xi_{00}^{-1}(1 - \phi_0) - \xi_{10}^{-1}(1 - \phi_1)
\]

which further simplifies to

\[
(\xi_{10} - \xi_{00})(1 - \theta) + (\xi_{00}^{-1} - \xi_{10}^{-1})(1 - \phi)
\]

under non differential misclassification. Note that (12) is a function of the misclassification probabilities and of the odds at \(Z = 0\). Consider the case of non differential misclassification with \(q, f \neq 1\). Then \(R_1 = 1\) implies that

\[
(\xi_{10} - \xi_{00}) \frac{1 - \theta}{\phi} + (\xi_{00}^{-1} - \xi_{10}^{-1}) \frac{1 - \phi}{\theta} \\
+ (\xi_{10} \xi_{00}^{-1} - \xi_{00} \xi_{10}^{-1}) \frac{1 - \phi}{\theta} \frac{1 - \theta}{\phi} = 0.
\]

It is easy to see that \(\xi_{10} > \xi_{00}\) implies that the LHS of (13) is positive. Likewise if \(\xi_{10} < \xi_{00}\) then the LHS of (13) is negative. Consequently the bias is zero if, and only if, \(\xi_{10} = \xi_{00}\), or equivalently if the odds ratio \(\xi_{10}/\xi_{00}\) is unity. Recall that \(b_1 = \log (\xi_{10}/\xi_{00})\). Therefore \(b_1 = 0\) implies that \(\gamma_1 = 0\) as well. This is the formal justification for the validity of the test of association on the observed data \(Y\) and \(X\) as a proxy for the relationship between \(Y\) and \(E\). The power of the test using the misclassified data is usually, but not always, diminished.

Next we calculate \(\Delta_2\) which is the bias induced in estimating \(\beta_2\), which is the regression parameter associated with \(Z\), when \(E\) is misclassified. From (7) and (9–10) we can show that

\[
\Delta_2(\theta, \phi) = \beta_2 - \gamma_2 = \log (R_2)
\]

where

\[
R_2 = \frac{[\xi_{01}(1 - \theta_0)/\phi_0, \xi_{10}(1 - \theta_1)/\phi_1]^{\otimes}}{[\xi_{00}(1 - \theta_0)/\phi_0, \xi_{11}(1 - \theta_1)/\phi_1]^{\otimes}}.
\]

Note that (14) involves all the odds \((\xi_{ik})\). The first order approximation for \(\Delta_2\) is

\[
(\xi_{01} - \xi_{00})(1 - \theta_0) + (\xi_{10} - \xi_{11})(1 - \theta_1)
\]

which interestingly does not involve \(\phi\). It is easy to see that \(R_2 = 1\) if

\[
(\xi_{10} - \xi_{11}) \frac{1 - \theta_1}{\phi_1} + (\xi_{01} - \xi_{00}) \frac{1 - \theta_0}{\phi_0} + (\xi_{10} \xi_{01} - \xi_{11} \xi_{00}) \\
\times \frac{1 - \theta_0}{\phi_0} \frac{1 - \theta_1}{\phi_1} = 0.
\]
Note that if (15) holds for all \( \theta \) and \( \phi \) then \( \xi_{10} = \xi_{11} \) and \( \xi_{00} = \xi_{01} \) must hold. This condition implies that the odds of the exposure \( E \) are independent of \( Z \) for both values of \( Y \). This condition also implies that \( \beta_3 = 0 \). However the equality \( \beta_3 = 0 \) does not generally zero (15). Hence the misclassification of the exposure \( E \) may lead to bias in estimating the effect of the covariate \( Z \) even if there is no interaction between them. Also note that under non differential misclassification the equalities \( \xi_{10} = \xi_{00} \), which is equivalent to \( \beta_1 = 0 \), and \( \xi_{11} = \xi_{01} \) also zero (15). Again the converse does not necessarily hold. It is also interesting to note that if \( \theta_i = 1 \), i.e., \( E = 1 \) is never misclassified then \( \Delta_2 = 0 \). Finally, using (8) and (9–10) we calculate the bias in estimating the interaction.

\[
\Delta_3(\theta, \phi) = \beta_3 - \gamma_3 = \log (R_3)
\]

where

\[
R_3 = \left[ \frac{\xi_{00}(1 - \theta_0)/\phi_0, \xi_{10}^{-1}(1 - \phi_1)/\theta_1, \xi_{01}^{-1}(1 - \phi_0)/\theta_0, \xi_{11}(1 - \theta_1)/\phi_1]}{\xi_{00}(1 - \theta_0)/\phi_0, \xi_{10}(1 - \theta_1)/\phi_1, \xi_{01}(1 - \phi_0)/\theta_0, \xi_{11}^{-1}(1 - \phi_1)/\theta_1} \right]^\circledast.
\] (16)

The first order approximation for (16) is

\[
(\xi_{00} - \xi_{01}) (1 - \theta_0) + (\xi_{11} - \xi_{10}) (1 - \theta_1) + (\xi_{01}^{-1} - \xi_{00}^{-1}) (1 - \phi_0) + (\xi_{10}^{-1} - \xi_{11}^{-1}) (1 - \phi_1).
\]

It is easily seen that \( R_3 = 1 \) under the same conditions as \( R_2 = 1 \).

### 3.2 Direction of the bias

We note that the biases \( \Delta_0 \) and \( \Delta_1 \) equal zero if \( \beta_1 = 0 \). In addition if \( \xi_{11} = \xi_{01} \) then \( \Delta_2 \) and \( \Delta_3 \) equal zero as well. Hence the error in estimation is not propagated if the misclassified covariate is not related with the outcome. A similar result is available for linear measurement error models (Carroll et al., 1995). More generally the quantities \( \Delta_2 \) and \( \Delta_3 \) are zeroed if the odds of the exposure \( E \) are independent of \( Z \) for both values of \( Y \). This condition is slightly stronger than \( \beta_3 = 0 \), or no interaction. The effect of the misclassification parameters on the direction of the bias is exhibited in Table 1.

<table>
<thead>
<tr>
<th>Change</th>
<th>( \Delta_0 )</th>
<th>( \Delta_1 )</th>
<th>( \Delta_2 )</th>
<th>( \Delta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_0 \uparrow )</td>
<td>( \downarrow )</td>
<td>( \uparrow )</td>
<td>( \uparrow ) if ( \xi_{00} &gt; \xi_{01} )</td>
<td>( \downarrow ) if ( \xi_{00} &gt; \xi_{01} )</td>
</tr>
<tr>
<td>( \theta_1 \uparrow )</td>
<td>( \uparrow )</td>
<td>( \downarrow )</td>
<td>( \uparrow ) if ( \xi_{11} &gt; \xi_{10} )</td>
<td>( \downarrow ) if ( \xi_{11} &gt; \xi_{10} )</td>
</tr>
<tr>
<td>( \phi_0 \uparrow )</td>
<td>( \uparrow )</td>
<td>( \downarrow )</td>
<td>( \uparrow ) if ( \xi_{00} &gt; \xi_{01} )</td>
<td>( \downarrow ) if ( \xi_{00} &gt; \xi_{01} )</td>
</tr>
<tr>
<td>( \phi_1 \uparrow )</td>
<td>( \downarrow )</td>
<td>( \uparrow )</td>
<td>( \uparrow ) if ( \xi_{11} &gt; \xi_{10} )</td>
<td>( \downarrow ) if ( \xi_{11} &gt; \xi_{10} )</td>
</tr>
</tbody>
</table>
3.3 The bias around a fixed value of the parameter

As pointed out by a referee expansions around a particular value of $\beta$ may also be of interest. For example, fixing $\theta$ and $\phi$ and assuming non differential misclassification we can expand $\Delta_k$ about $\beta_1$. In practice this is done by setting $\xi_{10} = \rho \xi_{00} + \delta$ and expanding (11), (12), (14) and (16) about $\delta = 0$. With some algebra we can show that for $\theta, \phi \neq 0$ the first order expansion is surprisingly simple, i.e.,

$$
\left[ \log \left( \xi_{00} + \frac{\phi}{1-\theta} \right) - \log \left( \rho \xi_{00} + \frac{\phi}{1-\theta} \right) \right] \\
+ \left( \rho \xi_{00} + \frac{\phi}{1-\theta} \right)^{-1} \delta + O(\delta^2) 
$$

(17)

for $k = 0, 1, 2, 3$. The first term of (17) reflects the bias when $\xi_{10}$ is correctly specified and $\theta, \phi$ are different than unity. Note that, as expected, this term vanishes when $\rho = 1$ (or $\beta_1 = 0$). Higher order terms reflect the impact of incorrect specification of the ratio $\xi_{10}/\xi_{00}$.

4. Numerical Example

We reconsider a case-control study on sudden infant death syndrome (SIDS) examined by Greenland (1988). The study investigated the relationship between SIDS and two covariates: gender and antibiotic use by the mother during pregnancy. Information regarding antibiotic use was gathered by interviewing the mother and is subject to misclassification. The relevant data can be found in Greenland’s Table 3. Greenland provides a corrected analysis based on an external validation study. For our purposes we do not take this external study into account but rather use the tools derived above to investigate the effect of various misclassification rates on the parameters of the logistic regression. The logistic regression model (4) was fit to the raw data resulting in:

$$
\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (-0.1641, -0.1757, 0.1487, 0.8355). 
$$

Table 2 presents the computed exact and approximate biases using equations (11), (12), (14), and (16), and assuming a range of hypothetical non-differential misclassification rates. For this example $\theta$ is the probability that the mother reports antibiotic use when it was taken while $\phi$ is the probability that the mother reports no use of antibiotics when in fact no antibiotics were used. Using a validation sample Greenland estimated $\hat{\theta} = 0.6024$ and $\hat{\phi} = 0.9014$. With these values we calculate $\Delta = (-0.034, 0.17, 0.15, -1.0)$ which amounts to substantial bias.

We see that if $\theta = 1$ only the coefficients relating to antibiotic exposure and its interaction with gender are biased. This of course is predicted by the theory. For
\[ \phi = 1, \Delta_0 = -\Delta_1 \text{ and } \Delta_2 = -\Delta_3. \] It can be seen from the bias formulae that this result is not restricted to this particular data set but holds in general. In general the bias increases with decreasing \( q \) and \( f \) (i.e. increasing misclassification rates). The biases can be substantial even for small misclassification rates and can be on the same or even larger order of magnitude than the uncorrected parameter estimates.

The biases are less sensitive to changes in \( \phi \) than changes in \( \theta \). The lowest value used for \( \phi \) is 0.862. When lower values are used negative values for some \( \pi \)'s are obtained. This implies that values of \( \phi \) lower than 0.862 are incompatible with the observed data. The large biases observed for values of \( \phi \) near its boundary are now easily explained. Since small values of \( \phi \) induce small values for some \( \pi \)'s they also imply either large (or small) values of the odds \( \xi_{ij} \), which with their inverses determine the bias. This also explains why the approximate formulas which are accurate for small misclassification rates behave poorly near their boundary value. For our example no restrictions were found on \( \theta \). In other data sets we may find higher sensitivity to changes in \( \theta \) rather than in \( \phi \). Consequently, for this example, in order to provide meaningful corrections it is more important to have an accurate knowledge of \( \phi \) than of \( \theta \). This information can be used in planning validation studies. Finally we note that the approximation (17) is less accurate for most values of the misclassification probabilities.
5. Other Misclassification Schemes

For the sake of completeness we briefly examine two additional misclassification schemes.

5.1 Covariate dependent misclassification of a exposure

We consider the situation where the exposure is misclassified. However, in contrast with our developments in Section 2 and 3, we assume that the misclassification probabilities depend on an additional covariate in the model and not the outcome. We assume that the model (1) holds and that $X$ is observed and is a misclassified version of $E$, the true value of the exposure. Define the misclassification probabilities

$$
\mu_i = \Pr [X = 1 \mid E = 1, Z = i],
$$
$$
\nu_i = \Pr [X = 0 \mid E = 0, Z = i].
$$

Note that these probabilities are a function of $Z$ the additional covariate in the model. The observed data is distributed according to $q_{ijk} = \Pr [Y = i, X = j, Z = k]$. Note that the distribution $q$ is a function of $p$ and the misclassification probabilities $\mu$ and $\nu$. Define

$$
\eta_{ij} = \frac{\Pr [E = 1 \mid Y = i, Z = j]}{\Pr [E = 0 \mid Y = i, Z = j]},
$$

the retrospective odds of exposure. This quantity is analogous to $\xi_{ij}$ of Section 3. Following the mathematics in Section 3 and omitting the details for brevity we can show that the induced biases, denoted by $\Upsilon_k$, are given by

$$
\Upsilon_0 = \log \left(\frac{1 + \eta_{00}(1 - \nu_0)/\mu_0}{1 + \eta_{10}(1 - \nu_0)/\mu_0}\right),
$$
$$
\Upsilon_1 = \log \left(\frac{[\eta_{00}(1 - \nu_0)/\mu_0, \eta_{10}(1 - \nu_0)/\mu_0]}{[\eta_{00}(1 - \nu_0)/\mu_0, \eta^{-1}_{10}(1 - \nu_0)/\mu_0]}\right),
$$
$$
\Upsilon_2 = \log \left(\frac{[\eta_{01}(1 - \nu_1)/\mu_1, \eta_{11}(1 - \nu_0)/\mu_0]}{[\eta_{00}(1 - \nu_0)/\mu_0, \eta_{11}(1 - \nu_1)/\mu_1]}\right),
$$

and

$$
\Upsilon_3 = \log \left(\frac{[\eta_{00}(1 - \nu_0)/\mu_0, \eta_{01}(1 - \nu_1)/\mu_1, \eta^{-1}_{10}(1 - \nu_0)/\mu_0, \eta_{11}(1 - \nu_1)/\mu_1]}{[\eta_{00}(1 - \nu_0)/\mu_0, \eta_{01}(1 - \nu_1)/\mu_1, \eta_{10}(1 - \nu_0)/\mu_0, \eta^{-1}_{11}(1 - \nu_1)/\mu_1]}\right).
$$
The first order approximations for these quantities are
\[
(1 - \nu_0) (\eta_{10} - \eta_{00}), \\
(1 - \nu_0) (\eta_{10} - \eta_{00} + \eta_{00}^{-1} - \eta_{10}^{-1}), \\
(1 - \nu_0) (\eta_{10} - \eta_{00}) + (1 - \nu_1) (\eta_{01} - \eta_{11}), \\
(1 - \nu_0) (\eta_{00} - \eta_{10} + \eta_{10}^{-1} - \eta_{00}^{-1}) + (1 - \nu_1) (\eta_{11} - \eta_{01} + \eta_{01}^{-1} - \eta_{11}^{-1}).
\]

Conditions for unbiasedness are immediate.

### 5.2 Covariate dependent misclassification of outcomes

Here we consider the situation where the outcomes (not a covariate) is misclassified and that the misclassification probabilities depend on the value of a covariate. Inference in this setting has been recently investigated by Nehaus (1999) and Cheng and Hesueh (1999). Suppose that we observe \( U \) a misclassified version of \( Y \) such that,
\[
\sigma_i = \Pr [U = 1 \mid Y = 1, E = i], \\
\tau_i = \Pr [U = 0 \mid Y = 0, E = i].
\]

Define
\[
\kappa_{ij} = \frac{\Pr [Y = 1 \mid E = i, Z = j]}{\Pr [Y = 0 \mid E = i, Z = j]}.
\]

Then the biases, \( \Psi_k \), in estimating the parameters may be expressed as
\[
\Psi_0 = \log \left( \frac{\tau_0}{\sigma_0} \frac{1 + \kappa_{00}^{-1}}{1 + \kappa_{00}^{-1} - \tau_0} \right), \\
\Psi_1 = \log \left( \frac{\sigma_0 \tau_1}{\sigma_1 \tau_0} \frac{\kappa_{00}^{-1}(1 - \tau_0)}{\kappa_{10}^{-1}(1 - \tau_1)} \frac{\kappa_{01}^{-1}(1 - \sigma_0)}{\kappa_{11}^{-1}(1 - \sigma_1)} \right), \\
\Psi_2 = \log \left( \frac{\kappa_{00}^{-1}(1 - \tau_0)}{\kappa_{00}(1 - \sigma_0)} \frac{\kappa_{01}^{-1}(1 - \tau_1)}{\kappa_{01}(1 - \sigma_1)} \right),
\]
and
\[
\Psi_3 = \log \left( \frac{\kappa_{00}(1 - \sigma_0)}{\kappa_{00}^{-1}(1 - \tau_0)} \frac{\kappa_{01}^{-1}(1 - \tau_0)}{\kappa_{01}(1 - \sigma_0)} \frac{\kappa_{10}^{-1}(1 - \tau_1)}{\kappa_{10}(1 - \sigma_1)} \frac{\kappa_{11}^{-1}(1 - \tau_1)}{\kappa_{11}(1 - \sigma_1)} \right).
\]
The first order approximations for these quantities are

\[
(1 - \sigma_0) \left(1 + \kappa_{00}\right) - (1 - \tau_0) \left(1 + \kappa_{00}^{-1}\right),
\]

\[
(1 - \tau_0) \left(1 + \kappa_{00}^{-1}\right) - (1 - \tau_1) \left(1 + \kappa_{10}^{-1}\right) + (1 - \sigma_0) \left(1 + \kappa_{01}\right)
\]

\[
- (1 - \sigma_0) \left(1 + \kappa_{00}\right),
\]

\[
(1 - \tau_0) \kappa_{00}^{-1} - (1 - \tau_1) \kappa_{01} + (1 - \sigma_0) \left(\kappa_{01} - \kappa_{00}\right),
\]

\[
(1 - \tau_0) \left(\kappa_{01}^{-1} - \kappa_{00}^{-1}\right) + (1 - \sigma_0) \left(\kappa_{00} - \kappa_{01}\right) + (1 - \tau_1) \left(\kappa_{10}^{-1} - \kappa_{11}^{-1}\right)
\]

\[
+ (1 - \sigma_0) \left(\kappa_{11} - \kappa_{10}\right).
\]

6. Summary and Discussion

In this paper we study the effect of misclassification on the parameters of a logistic regression when the covariates are binary. We focus on the differential misclassification of a covariate and derive exact and approximate formulas for the bias on the log-odds scale. Our calculations show that when the outcome and the exposure are related then misclassification may cause considerable bias for all the modeled covariates. For example suppose that there is no interaction between the exposure \(E\) and the covariate, i.e., \(\beta_3 = 0\). However if we observe only \(X\), the misclassified version of \(E\), we may decide that \(\gamma_3\) is significantly different than zero. This is important in practical data analysis because it means that in the presence of misclassification an additive logistic model, that is an unsaturated model, may look like a saturated one and visa versa. Throughout this paper the approximate formulas for the bias are derived by retaining the first term of the multivariate series (Taylor) expansion of \(\Delta_k\) about \(\theta = 1\) and \(\phi = 1\). These formulas hold for all values of \(\beta\). In addition approximate formulas for the bias for specific values of \(\beta\) and arbitrary \(\theta\) and \(\phi\) are derived.

Our motivation is to provide a tool to aid investigators in the design of both prospective and retrospective studies. Using our formulas one can assess the direction and magnitude of the bias (caused by misclassification) before the study is actually performed. For example the misclassification probabilities may be estimated using a small pilot study. If the resulting biases are significant than a method for correcting theses biases should be considered. Such methods are usually based on external information regarding the misclassification rates or an internal double sampling scheme or a validation study. Either way correcting for misclassification will require additional resources and the formulas developed may help investigators to decide if and when such corrections are warranted. Methods for such corrections are described by Reade-Christopher and Kupper (1991) when an auxiliary variable, which is correctly classified, is available. This method is closely related to double sampling methods (Tenenbein, 1970; Davidov and Haitovsky, 2000). Estimation procedures for misclassified data based on log-linear
models are summarized by Chen (1989). More recently Morrissey and Spiegelman (1999) review and compare several estimation methods for correcting for misclassification in $2 \times 2$ tables. Optimal designs for this problem have been addressed by Dahm et al. (1995) and Holcroft and Spiegelman (1999).

We consider two additional modes of misclassification, namely, covariate dependent misclassification of an exposure and covariate dependent misclassification of outcomes and derive exact and approximate formulas in both cases. The resulting formulas are similar to those obtained earlier but with different interpretation of the odds. We note that in all situations the number of categories for each exposure can also be increased and in this way approximate a continuous covariate. However an extension to true continuous covariates is more difficult because closed form solutions are unavailable in this case.

Acknowledgment

We thank the referees for their comments and suggestions which have improved this paper.

References


