Bias analysis for logistic regression with a misclassified multicategorical covariate

Juxin Liu$^1$ and Yaqing Liu$^2$

$^1$ Department of Mathematics and Statistics, University of Saskatchewan

$^2$ School of Public Health, University of Saskatchewan

April 28, 2011

Abstract

We extend the analysis of asymptotic biases in Davidov et. al. (2003) from a binary covariate to a multicategory covariate which is subject to misclassifications. The formulas are derived under the assumption of differential misclassification and can be simplified under the nondifferential misclassification. Further, provided with the assumption of nondifferential misclassification, we can determine the signs of the asymptotic biases under some certain circumstances. For a sensible understanding of how the bias may be influenced by the misclassification parameters, we use an illustrative example. It shows that the magnitude of bias increases as the frequency of misclassifications increases.

Keywords: misclassification; nondifferential misclassification; multicategorical covariate; logistic regression; odds ratio.

1 Introduction

In many epidemiological studies, the central interest is to find the association between a binary health outcome and a categorical exposure variable (or covariate). One common difficulty encountered in practice is that exposure variables are not always assessed correctly. For example, self-reported tobacco use has been widely used to measure smoking prevalence in epidemiological studies. However, some people may not honestly report
their smoking habit when being interviewed. Due to this possible error, the actual rel-
ationship between smoking and some certain disease (e.g. lung cancer) may not be
detected. Then this misleading result may influence the decision makers on allocating
resources for the prevention of smoking. The consequences of ignoring the errors in mea-
surement have been studied in the past over 40 years. For a textbook treatment of this
topic, please refer to Carroll et. al. (2006).

There have been extensive work focusing on the correction for the misclassification
of a binary exposure variable. To name a few, Greenland (1988) proposed a method for
variance estimation of the estimates under misclassifications; Davidov et. al. (2003) stud-
ied the asymptotic bias caused by a misclassified binary covariate in logistic regression
context; Prescott and Garthwaite (2005) developed a Bayesian method to account for
the misclassification in a binary risk factor under a prospective logistic regression model.
However, there has been very limited work on the misclassifications of a multcategory
covariate. In this paper, our focus is the misclassification in a multicategorical variable.
We extend the formulas derived in Davidov et. al. (2003) to a general setting where the
error-prone covariate has at least two different categories.

The remainder of this paper is organized as follows. Section 2 gives the details of
deriving the asymptotic biases due to the misclassifications of a multicategory covariate.
Via a data example, Section 3 illustrates how the biases vary across different values of
misclassification parameters. Section 4 concludes the findings and also discusses some
possibilities of the future work.

2 Derivation of the Asymptotic Bias

2.1 Notations

We consider a logistic regression model with a binary outcome, denoted by \(Y\), and a
categorical covariate which is subject to misclassifications. Let \(E\) denote the true value
of the categorical covariate and \(X\) denote the observed value of \(E\) (subject to errors).
Note that in practice, often \(E\) is not observable while \(X\) is observable. Suppose that
there are \(m(\geq 2)\) categories of \(E\), labeled by 1, 2, \ldots, \(m - 1\) and \(m\), respectively. Please
note those categories may or may not have orderings, that is, \(E\) can be either nominal
or ordinal. We define the following indicator variables for \(E\). For \(i = 1, 2, \ldots, m - 1\), let

\[
I_i = \begin{cases} 
1, & \text{if } E = i + 1, \\
0, & \text{otherwise}.
\end{cases}
\]

The logistic regression model that can describe the association between \(Y\) and \(E\) is

\[
\logit P(Y = 1|E) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \cdots + \beta_{m-1} I_{m-1}.
\]
Similarly, we can define the indicator variables for $X$. For $i = 1, 2, \ldots, m - 1$, let

$$I_i^* = \begin{cases} 1, & \text{if } X = i + 1, \\ 0, & \text{otherwise}. \end{cases}$$

Therefore, the logistic regression model for the observables $(Y, X)$ is

$$\text{logit} P(Y = 1|X) = \gamma_0 + \gamma_1 I_1^* + \gamma_2 I_2^* + \cdots + \gamma_{m-1} I_{m-1}^*. \quad (2)$$

Note the parameters $\gamma$ in (2) stand for the large sample limit of the regression coefficient estimates $\hat{\gamma}_i$ corresponding to the error-prone $X$.

Let $\Delta_i = \beta_i - \gamma_i, i = 0, 1, \ldots, m - 1$, that is, the asymptotic bias in each regression coefficient. Obviously the magnitude of $\Delta_i$ can be used to assess the harmful consequence of misclassifications. The objective of the following subsection is to derive the formula for the asymptotic bias. Prior to the derivation, we need to introduce some other notations as follows. Let

$$\pi_{di} = P(Y = d, E = i), \ d = 0, 1; \ i, j = 1, 2, \ldots, m,$$

$$p_{di} = P(Y = d, X = i),$$

$$\theta_{d,ij} = P(X = i|E = j, Y = d),$$

$$\xi_{d,ab} = \frac{\pi_{da}}{\pi_{db}}, \ a, b = 1, 2, \ldots, m.$$ Here $\theta_{d,ij}(i \neq j)$ stand for the misclassification rates. The diagonal elements ($i = j$) are correct classification rates, which are just sensitivity and specificity when $m = 2$. Obviously $\sum_{i=1}^{m} \theta_{d,ij} = 1$ for any given $d, j$. By the definition of conditional probability, we have

$$p_{di} = P(Y = d, X = i) = \sum_j P(X = i, E = j, Y = d)$$

$$= \sum_j P(X = i|E = j, Y = d)P(Y = d, E = j)$$

$$= \sum_j \theta_{d,ij} \pi_{dj}. \quad (3)$$

This can be rewritten in terms of matrices, that is,

$$p_d = \theta_d \pi_d, \quad (4)$$

where $p_d = (p_{d1}, p_{d2}, \ldots, p_{dm})'$, $\theta_d = (\theta_{d,ij}), \pi_d = (\pi_{d1}, \pi_{d2}, \ldots, \pi_{dm})'$. Note that the notation $\xi_{d,ab}$ is introduced to simplify the expression of asymptotic bias. The meaning of $\xi_{d,ab}$ can be better understood by rewriting it as $P(E = a|Y = d)/P(E = d|Y = d)$.
That is, the ratio of the exposure risk of categories $a$ and $b$ given $Y = d$. When the number of categories $m = 2$, $\xi_{d,ab}$ is simply the odds in favour of the category $a$. Moreover, the ratio of $\xi_{1,ab}/\xi_{0,ab}$ can be linked to the regression coefficients. According to the model (1), we have

$$\beta_0 = \logit P(Y = 1|E = 1),$$
$$\beta_0 + \beta_k = \logit P(Y = 1|E = k + 1), \ k = 1, 2, \ldots, m - 1.$$

Therefore,

$$e^{\beta_k} = \frac{\text{odds}(P(Y = 1|E = k + 1))}{\text{odds}(P(Y = 1|E = 1))} = \frac{P(E = k + 1|Y = 1)}{P(E = k + 1|Y = 0)} \cdot \frac{P(E = 1|Y = 1)}{P(E = 1|Y = 0)},$$
$$= \frac{\xi_{1,(k+1)1}}{\xi_{0,(k+1)1}}.$$

The second equality follows the definition of conditional probability. Therefore, $\beta_k$ is the logarithm of the ratio of the relative exposure risk between categories $k + 1$ and 1 in case group ($Y = 1$ for diseased) and the relative exposure risk in control group ($Y = 0$ for undiseased). Particularly when $m = 2$, $\beta_1$ is the logarithm of the odds ratio of exposure risk, i.e., $\text{odds}(P(E = 2|Y = 1))/\text{odds}(P(E = 2|Y = 0))$. Similarly, replacing $E = 1$ by $E = l + 1$ in the above derivation, we have

$$e^{\beta_k - \beta_l} = \frac{\xi_{1,(k+1)(l+1)}}{\xi_{0,(k+1)(l+1)}}, \ l \geq 1, \ k \geq 1.$$

That is, $\beta_k - \beta_l$ is the logarithm of the ratio of the relative exposure risk between categories $k + 1$ and $l + 1$ in case group and the relative exposure risk in control group.

### 2.2 Asymptotic Bias

In this subsection, we derive the expressions of the asymptotic biases $\Delta_i$ as functions of the misclassification rates. Let $\pi_{d|c} = P(Y = d|E = c)$ and $p_{d|c} = P(Y = d|X = c), d = 0, 1, c = 1, 2, \ldots, m$. According to the models (1) and (2), we have

$$\Delta_0 = \beta_0 - \gamma_0 = \logit(\pi_{1|1}) - \logit(p_{1|1}) = \log\left(\frac{\pi_{1|1}}{\pi_{0|1}}\right) - \log\left(\frac{p_{1|1}}{p_{0|1}}\right).$$

The second equality is due to the facts that

$$P(Y = 0|E) + P(Y = 1|E) = 1,$$
$$P(Y = 0|X) + P(Y = 1|X) = 1.$$
By a straightforward use of conditional probability definition, the above can be re-written as

\[ \Delta_0 = \log \left( \frac{\pi_{11}}{\pi_{01}} \right) - \log \left( \frac{p_{11}}{p_{01}} \right) \]

Thanks to the formula (3) that expresses \( p_{di} \) in terms of \( \pi_{di} \) and \( \theta_{d,ij} \), we have

\[ \Delta_0 = \log \left( \frac{\pi_{11} \sum_{j=1}^{m} \theta_{0,1j} \pi_{0j}}{\pi_{01} \sum_{j=1}^{m} \theta_{1,1j} \pi_{1j}} \right) = \log \left( \frac{\sum_{j=1}^{m} \theta_{0,1j} \xi_{0,1j}}{\sum_{j=1}^{m} \theta_{1,1j} \xi_{1,1j}} \right). \]

Similarly, we can find the expressions of \( \Delta_c = \beta_c - \gamma_c \), \( c = 1, \ldots, m - 1 \). The models (1) and (2) imply that \( \logit(P(Y = 1|E = c + 1)) = \beta_0 + \beta_c \) and \( \logit(P(Y = 1|X = c + 1)) = \gamma_0 + \gamma_c \). Therefore, we have

\[ \Delta_c = \beta_c - \gamma_c = \logit(\pi_{1|c+1}) - \logit(p_{1|c+1}) - (\beta_0 - \gamma_0) = \log \left( \frac{\sum_{j=1}^{m} \theta_{0,(c+1)j} \xi_{0,(c+1)j}}{\sum_{j=1}^{m} \theta_{1,(c+1)j} \xi_{1,(c+1)j}} \right) - \Delta_0. \]

Since the asymptotic biases can be viewed as a function of \( \theta_{d,ij} \), we consider an approximate form of the asymptotic biases by first-order Taylor expansion in the same spirit of Davido et al (2003). It is worth noting that the Taylor expansion is around the particular misclassification matrix \( \theta_d = I_m \), which stands for no misclassifications. Thus, we have the following approximations:

\[ \Delta_0 \approx \sum_{j=1}^{m} (\theta_{0,1j} \xi_{0,1j} - \theta_{1,1j} \xi_{1,1j}), \]

\[ \Delta_c \approx \sum_{j=1}^{m} \left( \theta_{0,(c+1)j} \xi_{0,(c+1)j} - \theta_{1,(c+1)j} \xi_{1,(c+1)j} \right) - \sum_{j=1}^{m} (\theta_{0,1j} \xi_{0,1j} - \theta_{1,1j} \xi_{1,1j}), c = 1, 2, \ldots, m - 1. \]

### 2.3 Nondifferential Misclassification

The derivation in the previous subsection is based on differential misclassification. That is, the misclassification parameters are different for different values of \( Y \). In this subsection, we focus on nondifferential misclassification. Therefore, the misclassification
parameters are the same for different values of $Y$. Hence, we can suppress the first subscript in $\theta_{d,ij}$, i.e., $\theta_{ij}$. Now the asymptotic bias $\Delta_0$ can be simplified as below.

$$\Delta_0 = \log \left( \frac{\sum_{j=1}^{m} \theta_{1j} \xi_{0,j1}}{\sum_{j=1}^{m} \theta_{1j} \xi_{1,j1}} \right)$$

$$= \log \left( \frac{\sum_{j=1}^{m} \theta_{1j} \xi_{0,j1}}{\sum_{j=1}^{m} \theta_{1j} \xi_{1,j1}} \right)$$

$$= \log \left( \frac{\theta_{11} + \sum_{j>1} \theta_{1j} \xi_{0,j1}}{\theta_{11} + \sum_{j>1} \theta_{1j} \exp(\beta_{j-1}) \xi_{0,j1}} \right).$$

Obviously, if $\beta_k = 0, k = 1, \ldots, m - 1$, then $\Delta_0 = 0$, that is, no asymptotic bias. Analogously, we can simplify the expressions of $\Delta_c, c = 1, \ldots, m - 1$ as well:

$$\Delta_c = \log \left( \frac{\theta_{(c+1)(c+1)} + \sum_{j \neq c+1} \theta_{(c+1)j} \xi_{0,j(c+1)}}{\theta_{(c+1)(c+1)} + \sum_{j \neq c+1} \theta_{(c+1)j} \exp(\beta_{j-1}^* - \beta_{c}) \xi_{0,j(c+1)}} \right) - \Delta_0,$$

where $\beta_{j-1}^* = \beta_{j-1}$ if $j > 1$ and zero otherwise. Therefore, if $\beta_{j-1} = \beta_{c} = 0$, i.e., all regression coefficients are zero except for the intercept, there are no biases($\Delta_c = 0$). In summary, if $\beta_1 = \beta_2 = \ldots = \beta_{m-1}$, then the asymptotic biases are all zero with the assumption of nondifferential misclassification.

Moreover, we can find the direction of the bias under some certain circumstances. For each $k \geq 1$, if $\beta_k > 0$ then $\Delta_0 < 0$. In other words, $\hat{\gamma}_0$, the intercept estimator based on $X$, is (asymptotically) inflated. Otherwise if all $\beta_k (k \geq 1)$ are negative, then $\Delta_0 > 0$. Let $\beta_{-c} = \{\beta_1, \beta_2, \ldots, \beta_{c-1}, \beta_{c+1}, \ldots, \beta_{m-1}\}$, i.e., the collection of all regression coefficients except for $\beta_c$ and $\beta_0$. For a given category $c+1$, if $\beta_c < \min\{0, \beta_{-c}\}$, then the first item in $\Delta_c$ is negative. Notice the fact that $\Delta_0 > 0$ when $\beta_k (k \geq 1)$ are all negative. Therefore, if $\beta_c < \min\beta_{-c}$ and $\beta_{-c} < 0$, then $\Delta_c < 0$. As an analogy, if $\beta_c > \max\beta_{-c}$ and $\beta_{-c} > 0$, then $\Delta_c > 0$. Nonetheless, the conditions on $\beta$ are sufficient but not necessary. A more general conclusion about the signs of biases needs to be investigated in the future.

### 3 An Illustrative Example

In this section, we use a data example to illustrate how the asymptotic bias would be influenced by the misclassification rates. The data is from Table 7.5 in Agresti (2002). This data example is about mental health for adult residents of Alachua County, Florida. Different from the regression task discussed in Agresti (2002), the objective here is to find the relationship between socioeconomic status ($Y$) and mental health status ($X$). In this
data example, $Y$ is measured as binary with 1 for high socioeconomic status and 0 for low socioeconomic status. The mental health status is an ordinal variable with four categories (well, mild, moderate, impaired). The sample size $n$ is 40. We did logistic regression for this data by introducing three indicators $I^*_{i} = 1$ if $X = i+1$ and zero otherwise, $i = 1, 2, 3$. Note that 1 stands for well, 2 stands for mild, 3 stands for moderate, and 4 stands for impaired. The regression coefficient for $I^*_2$ is -0.38 while the regression coefficient for $I^*_3$ is -0.22. This implies that the adults whose mental impairment is more severe tend to have higher chance of having high socioeconomic status, which certainly conflicts with the common sense. The conflict may be caused by potential misclassifications in mental health status. Note that the “mental impairment” is purely a mental construct and is subjectively evaluated at the likert scaling. Therefore, it is very likely that the measurements of mental impairment are subject to errors. For example, people who have mild mental impairment can be identified as moderate mental impairment, and vice versa.

From the previous section, we have seen that asymptotic bias in regression coefficients can be expressed as a function of misclassification rates. Thanks to (4), the $\pi_d$ can be expressed as $\theta^{-1}_d p_d$. Thus based on the observed frequency $p_d$ for data $(x_i, y_i)$, $i = 1, \ldots, n$, we can derive the true cell probabilities $\pi_{di}$ and thus $\xi_{d,ab}$. In summary, we can evaluate the asymptotic biases based on the postulated misclassification matrices $\theta_d$ and observed $p_d$, $d = 0, 1$. Here we consider nondifferential misclassification, that is, $\theta_0 = \theta_1$. Due to the natural ordering among the four categories of mental health status, we assume that misclassifications only happen between two adjacent categories with equal probability, i.e., $(1 - \theta_{ii})/2$. For the category that has only 1 adjacent category such as category 1, the misclassification rate is simply 1 minus the correct classification rate, i.e., $1 - \theta_{ii}$. We consider four different misclassification mechanisms in an increasing order of the frequency of misclassifications. See Table 1 for the results. It shows that the magnitude of asymptotic bias increases when the amount of error increases.

To check how sensitive the bias is to the misclassifications associated with different categories, we consider another set of four different types of misclassification matrices. The $i$-th type stands for the case that most of the misclassifications is related to category $i$. Take $i = 1$ for an instance, the first type of misclassification matrix implies that the misclassification happens most frequently when the true mental health status is “well”. See Table 2 for the comparison. By comparing the length of the bias vector, it turns out the bias is most sensitive to the misclassification associated with category 4. In other words, the correct classification of category 4 is the most important in the sense of reducing the bias.

A worth noting fact is the substantial difference between the exact and approximate evaluations of the biases when the amount of errors is considerably large. It is not surprising because the Taylor approximation is not supposed to behave well when the misclassification parameters are far away from $I_m$ (stands for no misclassifications) at
which the Taylor series is centered. Therefore, examining the difference between the exact and approximate evaluations of biases can help check whether the misclassification is an issue. The larger the difference is, the more severe misclassification issue it indicates.

4 Discussion

We have derived the formulas of asymptotic biases in the regression coefficients when the covariate is subject to misclassifications in a logistic regression model. Different from the binary covariate considered in Davidov et. al. (2003), our work focuses on a multicategory covariate and thus includes a binary covariate as a special case. Similar to the idea in Davidov et. al. (2003), we also derive the first-order Taylor expansions of the asymptotic biases. Under a nondifferential misclassification assumption, the expressions of asymptotic biases can be simplified and thus more intuitive. To examine how the bias would be influenced by the misclassification parameters, we introduce one real data example and devise several different mechanisms for misclassification.

There are several issues that need to be further investigated in the future. First, a more general investigation of the direction of asymptotic biases is required. As indicated in Section 2.3, we have discovered the direction of the biases under some particular situations. However, a general determination of the sign of the biases is still an open problem. Second, error free covariate may need to be included in some research studies. As concluded from Daviov et. al., the regression coefficient for the error-free covariate $Z$ can also be biased due to the misclassification in the exposure variable $X$. Third, an estimation method is required for a finite sample. This article concerns the asymptotic biases due to the misclassified multicategorical $X$ as the sample size approaches infinity. In practice, the statistical inference based on a finite sample is often of interest as well. When performing the estimation procedure, one cautionary note is the underlying constraints of the misclassification parameters. For a binary covariate, it is well known that a constraint on misclassification parameters is necessary to ensure the model identifiability, i.e., sensitivity $> 1$–specificity (Gustafson 2004). Now for a multicategory covariate, the constraints on $\theta_{d,ij}$ should be certainly more complicated. The constraints may militate on implementing the maximum likelihood method. If there is a priori knowledge about the misclassification parameters, one may consider a Bayesian approach. Therefore, the constraints and the prior information should both be incorporated in the construction of priors.
Acknowledgments

This research has been supported by NSERC Canada Discovery Grants, New Faculty Graduate Student Support Program from the University of Saskatchewan.

References


Table 1: The data example about mental health: Exact and approximate biases for four types of misclassification parameters in an increasing of the amount of errors.

<table>
<thead>
<tr>
<th>Asymptotic Bias</th>
<th>Type I: $\theta_{ii}=(0.9, 0.9, 0.9, 0.9)$</th>
<th>Type II: $\theta_{ii}=(0.9, 0.9, 0.65, 0.9)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Approx</td>
</tr>
<tr>
<td>$\Delta_0$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\Delta_1$</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>$\Delta_2$</td>
<td>-0.38</td>
<td>-0.49</td>
</tr>
<tr>
<td>$\Delta_3$</td>
<td>0.04</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Asymptotic Bias</th>
<th>Type III: $\theta_{ii}=(0.65, 0.9, 0.65, 0.9)$</th>
<th>Type IV: $\theta_{ii}=(0.65, 0.6, 0.6, 0.9)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Approx</td>
</tr>
<tr>
<td>$\Delta_0$</td>
<td>-0.02</td>
<td>-0.01</td>
</tr>
<tr>
<td>$\Delta_1$</td>
<td>0.99</td>
<td>3.51</td>
</tr>
<tr>
<td>$\Delta_2$</td>
<td>-0.24</td>
<td>-0.20</td>
</tr>
<tr>
<td>$\Delta_3$</td>
<td>0.20</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 2: The data example about mental health: Exact and approximate biases for four types of misclassification parameters. The $i$-th type stands for the situation that most of the misclassification is related to category $i$.

<table>
<thead>
<tr>
<th>Asymptotic Bias</th>
<th>Type 1: $\theta_{ii}=(0.75, 0.9, 0.9, 0.9)$</th>
<th>Type 2: $\theta_{ii}=(0.9, 0.75, 0.9, 0.9)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Approx</td>
</tr>
<tr>
<td>$\Delta_0$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\Delta_1$</td>
<td>0.08</td>
<td>0.12</td>
</tr>
<tr>
<td>$\Delta_2$</td>
<td>-0.31</td>
<td>-0.38</td>
</tr>
<tr>
<td>$\Delta_3$</td>
<td>0.03</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Asymptotic Bias</th>
<th>Type 3: $\theta_{ii}=(0.9, 0.9, 0.75, 0.9)$</th>
<th>Type 4: $\theta_{ii}=(0.9, 0.9, 0.9, 0.75)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Approx</td>
</tr>
<tr>
<td>$\Delta_0$</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>$\Delta_1$</td>
<td>0.21</td>
<td>0.24</td>
</tr>
<tr>
<td>$\Delta_2$</td>
<td>-0.37</td>
<td>-0.39</td>
</tr>
<tr>
<td>$\Delta_3$</td>
<td>0.12</td>
<td>0.11</td>
</tr>
</tbody>
</table>