A Bayesian Analysis of Binary Misclassification

Christopher R. Bollinger[a], Martijn van Hasselt[b]

[a] *Department of Economics, The University of Kentucky*
[b] *Department of Economics, The University of North Carolina Greensboro*

Abstract

We consider Bayesian inference about the mean of a binary variable that is subject to misclassification error. For several intuitive priors on the misclassification probabilities, we derive new analytical expressions for the posterior.
Key words: Bayesian inference, partial identification, misclassificiation

1. Introduction

We consider the problem of inference for the population mean of a binary variable that suffers from classification error. This type of model has a long history in both statistics and economics (e.g. **???**). If the misclassification rates are known or consistent estimates of the misclassification rates are available, the mean is identified and can be consistently estimated (**?**). In the absence of identification, only *partially* identification (a set of bounds) is available. The collection of feasible parameter values is called the identified set. In the classical approach to inference (e.g. **???**), a confidence interval for the parameter takes the form of the estimated bounds, plus a multiple of their standard errors. The resulting region in the parameter space, however, can be quite wide and classical inference provides no additional information about the location of the parameter within the bounds. However, in classical analysis there is no formal way to incorporate prior informal knowledge of the misclassification.

In this paper we take a Bayesian approach to inference for this problem. Our analysis relies on key insights of **?** and **?**. Some studies previously analyzed Bayesian models of misclassification but these achieved identification through the prior (**????**). In contrast, we consider a variety of priors that explicitly incorporate the parameter bounds inherent in the model. These priors can be considered intermediate between weak information leading only to partial identification and strong information leading to full identification. Moreover, we derive exact, analytical expression for the posterior and do not rely on Markov chain Monte Carlo sampling.

Our approach also allows researchers to examine sensitivity of estimates with respect to assumptions about misclassification rates. The analysis examines how additional prior information about these rates affects what the researcher can learn about the population mean. The identification problem is by no means eliminated through the use of a Bayesian prior.The prior allows researchers to incorporate varying amounts of information and examine the effect on posterior inferences. Our results show that typical resulting posteriors are far from uniform and thus provide additional information about the location of the population mean within the identified set.

The remainder of this paper is organized as follows: section 2 introduces the model and discusses the limiting behavior of the posterior distribution of the unidentified parameters. Section 3 introduces several prior distributions that range from less to more informative about the probabilities of misclassification errors. The resulting finite-sample posterior distributions are presented in section 4. Section 5 provides concluding remarks. Derivations of some of the results are collected in the appendix.

2. A Bayesian Misclassification Model

Let $Z \in \{0, 1\}$ be a binary random variable with $P(Z = 1) = \pi$. Instead of observing Z, we observe $X \in \{0, 1\}$, which may suffer from misclassification error:

$$P(X = 1|Z) = p(1 - Z) + (1 - q)Z. \tag{1}$$

Here, p is the probability of a false positive, whereas q is the probability of a false negative. Further, let μ $\pi(1 - q) + (1 - \pi)p$ be the mean of X. As is typical in the literature, we assume that $p + q < 1$. This ensures that the misclassication is not so bad as to either reverse the categorical definitions ($Cov(X, Z) < 0$) or to sever any relationship between X and Z ($Cov(X, Z) = 0$). Using the equation for μ, we find

$$0 \leq p \leq \mu, \quad 0 \leq q \leq 1 - \mu. \tag{2}$$

The parameter π, however, can take values over the entire unit interval. For example, if $p = \mu$, then $\pi = 0$, regardless of the value of q. Similarly, if $q = 1-\mu$, then $\pi = 1$. Hence, π is completely unidentified. One might expect, then, that its posterior distribution will resemble a uniform distribution, but we will see shortly that this is not the case.

Given a random sample $\mathbf{X} = (X_1, \ldots, X_n)$, let $n_1 = \sum_{i=1}^{n} X_i$ and $n_0 = n - n_1$ be the observed number of ones and zeroes, respectively. We focus on a parameterization in terms of π and μ,so that the likelihood can be represented as

$$\begin{aligned} f(\mathbf{X}|\pi,\mu) &= \mu^{n_1}(1-\mu)^{n_0} \\ &= f(\mathbf{X}|\mu). \end{aligned}$$

This highlights the fact that μ is identified through the likelihood, whereas π is not. The joint posterior of π and μ is then

$$\begin{aligned} f(\pi,\mu|\mathbf{X}) &\propto f(\mathbf{X}|\pi,\mu)\cdot f(\pi,\mu) \\ &\propto f(\mathbf{X}|\mu)\cdot f(\mu)\cdot f(\pi|\mu) \\ &\propto f(\mu|\mathbf{X})\cdot f(\pi|\mu). \end{aligned}$$

The joint posterior of (π,μ) factors into the product of the marginal posterior of the identified parameter μ and the conditional prior of the unidentified parameter π given μ (Poirier 1998; Moon and Schorfheide 2012). In our context, it is now clear that the updating of beliefs occurs because the data are informative about μ. An important part of the joint posterior, however, is the conditional prior of the unidentified parameter.

If the true value of the population mean of X is μ_0, then under standard regularity conditions the posterior distribution of μ will increasingly concentrate around μ_0) as $n\to\infty$ (e.g. **??**). This has an important implication for the posterior of π. Since

$$\begin{aligned} f(\pi|\mathbf{X}) &= \int f(\pi,\mu|\mathbf{X})d\mu \\ &\propto \int f(\mu|\mathbf{X})f(\pi|\mu)d\mu, \end{aligned}$$

the marginal posterior of π can be seen as a continuous mixture of conditional priors, with mixing distribution $f(\mu|\mathbf{X})$. As the sample size increases, however, the mixing distribution becomes asymptotically degenerate at $\mu=\mu_0$

and $f(\pi|\mathbf{X})$ will converge to $f(\pi|\mu_0)$.[1] As such, in large samples the marginal posterior of π will be close to the conditional prior of π, evaluated at a particular value of μ.

3. Prior Distributions

In this section we present a number of natural prior distributions and their implied distributions of π given μ (which can be viewed as the asymptotic posterior of π). The priors are ordered in the sense that they represent situations with increasing amounts of prior information about the misclassification rates. The resulting finite-sample posterior distributions are given and further investigated in section 4.

A first candidate for a prior distribution is a uniform distribution for π, combined with a uniform distribution for p and q in the region where $p + q < 1$:

$$f_1(\pi, p, q) = 2 \cdot \mathbf{1}\{0 \leq \pi \leq 1, p + q < 1\}. \tag{3}$$

Here, $\mathbf{1}(\cdot)$ is the indicator function. With this choice of prior, it is shown in the appendix that

$$f_1(\mu) = -2\left[\mu \log \mu + (1-\mu)\log(1-\mu)\right],$$

which has an inverted u-shape with a mode at $\mu = 1/2$. The conditional prior of π is

$$f_1(\pi|\mu) = [f_1(\mu)]^{-1}\left[\frac{2(1-\mu)}{(1-\pi)}\mathbf{1}\{\pi < \mu\} + \frac{2\mu}{\pi}\mathbf{1}\{\pi \geq \mu\}\right], \tag{4}$$

[1] The argument given here also applies to the unidentified parameter p and q. Thus, in large samples $f(p|\mathbf{X})$ and $f(q|\mathbf{X})$ will converge to $f(p|\mu_0)$ and $f(q|\mu_0)$, respectively.

which has a unique mode at $\pi = \mu$.

The second prior we consider is a uniform prior for μ, combined with conditional priors $p|\mu \sim U(0, \mu)$ and $q|\mu \sim U(0, 1-\mu)$ that are uniform in the identified set:

$$f_2(\mu, p, q) = \frac{1}{\mu(1-\mu)} \mathbf{1}\{0 \leq p \leq \mu, 0 \leq q \leq 1-\mu\}. \tag{5}$$

It follows that

$$f_2(\mu, p, \pi) = \frac{\mu - p}{\mu(1-\mu)\pi^2}.$$

To calculate the prior of μ and π, it is necessary to consider the feasible values of p, given that q ranges from 0 to $1-\mu$. Using equation (**??**) it can be shown that

$$\max\left\{0, \frac{\mu - \pi}{1 - \pi}\right\} \leq p \leq \mu,$$

so that

$$\begin{aligned} f_2(\mu, \pi) &= \mathbf{1}\{\pi > \mu\} \int_0^{\mu} \frac{(\mu - p)}{\mu(1-\mu)\pi^2} dp \\ &\quad + \mathbf{1}\{\pi \leq \mu\} \int_{(\mu-\pi)/(1-\pi)}^{\mu} \int_0^{\mu} \frac{(\mu - p)}{\mu(1-\mu)\pi^2} dp \\ &= \mathbf{1}\{\pi > \mu\} \frac{\mu}{2(1-\mu)\pi^2} + \mathbf{1}\{\pi \leq \mu\} \frac{(1-\mu)}{2\mu(1-\pi)^2} = f_2(\pi|\mu) \end{aligned} \tag{6}$$

where the last equality follows because $f_2(\mu) = 1$.

The third prior distribution we consider expresses the belief that, conditional on μ, lower misclassification rates are more likely than higher ones. We combine a uniform prior for μ with 'power-type' conditional priors for p and q (which are proportional to $p^{-1/2}$ and $q^{-1/2}$). This yields the prior

$$f_3(\mu, p, q) = \frac{1}{4\sqrt{\mu(1-\mu)pq}} \mathbf{1}\{0 \leq p \leq \mu, 0 \leq q \leq 1-\mu\}. \tag{7}$$

This implies the following joint prior distribution for (μ, p, π):

$$f_3(\mu, p, \pi) = \frac{1}{4\pi\sqrt{\pi\mu(1-\mu)}} \cdot \frac{\mu - p}{\sqrt{p^2(1-\pi) + p(\pi - \mu)}},$$

where $\max\{0, (\mu - \pi)/(1-\pi)\} \le p \le \mu$. It is shown in the appendix that

$$f_3(\pi|\mu) = \frac{\mu(1-\pi) + \frac{1}{2}(\pi - \mu)}{4\pi(1-\pi)\sqrt{\pi(1-\pi)\mu(1-\mu)}} \log\left(\frac{\pi-\mu+2(1-\pi)\mu+2\sqrt{\pi(1-\pi)\mu(1-\mu)}}{|\pi-\mu|}\right) - \frac{1}{4\pi(1-\pi)}. \tag{8}$$

The fourth and final prior we consider expresses the belief that there is a probability λ that the misclassification error is *symmetric*. In that case, $p = q$ and false positive and false negatives are equally likely. Symmetry is a substantive assumption that changes the identified set. Similar to the asymmetric case, we maintain the assumption that Z and X are positively correlated, so that $p < \frac{1}{2}$. From the equation $\mu = (1-\pi)p + \pi(1-p)$, the parameter π can now be bounded in a non-trivial way:

$$\begin{array}{ll} 0 \le \pi \le \mu & \text{if} \quad \mu < \frac{1}{2} \\ \mu \le \pi \le 1 & \text{if} \quad \mu > \frac{1}{2} \end{array}. \tag{9}$$

The symmetry assumption is therefore strong in terms of shrinking the identified set[2], and the inequalities in (2) imply that $p \le \min\{\mu, 1-\mu\}$. A conditional prior that imposes the restriction $p = q$ and is uniform over the set of feasible values can be written as

$$\tilde{f}(p, q|\mu) = \frac{1}{\mu}\mathbf{1}\{p = q, p \le \mu < \tfrac{1}{2}\} + \frac{1}{1-\mu}\mathbf{1}\{p = q, \frac{1}{2} < \mu \le 1 - p\}. \tag{10}$$

[2]Since $\pi = (\mu - p)/(1 - 2p)$, it follows that π is identified and equal to $\frac{1}{2}$ if $\mu = \frac{1}{2}$. Identification of π in this case hinges on $p < \frac{1}{2}$ being a strict inequality. If $p = \frac{1}{2}$ is allowed, this result breaks down.

Using a uniform prior for μ, the mixture prior is

$$f_4(\mu,p,q) = \frac{\lambda}{\mu}\mathbf{1}\{p=q, p\le\mu<\tfrac{1}{2}\} + \frac{\lambda}{1-\mu}\mathbf{1}\{p=q, \frac{1}{2}<\mu\le 1-p\} + \frac{(1-\lambda)}{\mu(1-\mu)}\mathbf{1}\{0\le p\le\mu, 0\le q\le 1-\mu\}. \tag{11}$$

Thus, with probability λ the misclassification error is believed to be symmetric and p has a uniform distribution over the set of feasible values, and with probability $(1-\lambda)$ the error is believed to be asymmetric. Using a change of variables to (π,μ), it can be shown that

$$f_4(\pi|\mu) = \frac{\lambda(1-2\mu)}{\mu(1-2\pi)^2}\mathbf{1}\{\pi\le\mu<\tfrac{1}{2}\} + \frac{\lambda(2\mu-1)}{(1-\mu)(1-2\pi)^2}\mathbf{1}\{\tfrac{1}{2}<\mu\le\pi\} + \frac{(1-\lambda)\mu}{2(1-\mu)\pi^2}\mathbf{1}\{\pi>\mu\} + \frac{(1-\lambda)(1-\mu)}{2\mu(1-\pi)^2}\mathbf{1}\{\pi\le\mu\}. \tag{12}$$

4. Posterior Distributions

We now present analytical results for the finite-sample posterior distributions of π, using the priors discussed in the previous section. Derivations can be found in the appendix. The posteriors corresponding to $f_1(\pi,p,q)$ in equation (3) and $f_2(\mu,p,q)$ in equation (5) are given by

$$f_1(\pi|\mathbf{X}) \propto \frac{2}{\pi}\left[\left(\frac{n_0+1}{n+2}\right)\frac{\pi}{(1-\pi)} + \frac{\left(\frac{n_1+1}{n+2}-\pi\right)}{(1-\pi)}I_{n_1+1,n_0+1}(\pi)\right] - \frac{2}{(n+2)}\frac{1}{B_{n_1+1,n_0+1}}\pi^{n_1}(1-\pi)^{n_0}, \tag{13}$$

$$f_2(\pi|\mathbf{X}) = \frac{1}{2\pi^2}\left(\frac{n_1+1}{n_0}\right)I_{n_1+2,n_0}(\pi) + \frac{1}{2(1-\pi)^2}\left(\frac{n_0+1}{n_1}\right)(1-I_{n_1,n_0+2}(\pi)), \tag{14}$$

where $B_{a,b}$ is the Beta function and $I_{a,b}(t)$ is the cumulative distribution function of the Beta distribution with parameters (a,b). [3] For prior $f_3(\mu,p,q)$ in equation (7), the posterior of π is

$$f_3(\pi|\mathbf{X}) = \frac{1}{B_{n_1+1,n_0+1}} \int_0^1 \mu^{n_1}(1-\mu)^{n_0} f_3(\pi|\mu)d\mu, \tag{15}$$

where $f_3(\pi|\mu)$ is given in equation (8). This expression cannot be further simplified and we use numerical integration to calculate the posterior. Finally, using the mixture prior in equation (11), the corresponding posterior of π is a mixture distribution

$$f_4(\pi|\mathbf{X}) = \lambda\tilde{f}(\pi|\mathbf{X}) + (1-\lambda)f_2(\pi|\mathbf{X}), \tag{16}$$

where $f_2(\pi|\mathbf{X})$ is the posterior in equation (14), $\tilde{f}(\pi|\mathbf{X})$ is given by

$$\tilde{f}(\pi|\mathbf{X}) = \begin{cases} \frac{1}{(1-2\pi)^2}\left[\binom{n+1}{n_1} I_{n_1,n_0+1}(\pi,\tfrac{1}{2}) - 2I_{n_1+1,n_0+1}(\pi,\tfrac{1}{2})\right] & \text{if } \pi < \frac{1}{2} \\ \frac{1}{(1-2\pi)^2}\left[2\binom{n_1+1}{n_0} I_{n_1+2,n_0}(\tfrac{1}{2},\pi) - \binom{n+1}{n_0} I_{n_1+1,n_0}(\tfrac{1}{2},\pi)\right] & \text{if } \pi > \frac{1}{2} \end{cases},$$

and $I_{a,b}(s,t) = I_{a,b}(t) - I_{a,b}(s)$.

Graphs of the posteriors $f_1(\pi|\mathbf{X})$, $f_2(\pi|\mathbf{X})$ and $f_3(\pi|\mathbf{X})$ are given in figures 1, 2 and 3. We plot the finite-sample posteriors for sample sizes $n = 20$ and $n = 100$, when the observed fraction of ones is $n_1/n = 0.25$, as well as the conditional priors of π given μ, evaluated at $\mu = 0.25$. The latter represent the asymptotic limits of $f_j(\pi|X)$, $j = 1,2,3$ when $n \to \infty$ and the true value of μ is $\mu_0 = 0.25$. Figures 1 and 2 show that the priors in equations (3) and (5) lead to similar posteriors. The posteriors are informative in the sense that

[3] The function $I_{\alpha,\beta}(t)$ is also referred to as the regularized Beta function (Abramowitz and Stegun 1964).

they place higher probability on values of π that are close to 0.25 and lower probability on values close to 0 or 1. Figure 3 shows that under the more informative prior in equation (7), the posterior becomes more concentrated around $n_1/n = 0.25$.

[Figure 1 about here]

[Caption: Finite-sample posteriors $f_1(\pi|\mathbf{X})$ and conditional prior $f_1(\pi|\mu = 0.25)$]

[Figure 2 about here]

[Caption for Figure 2: Finite-sample posteriors $f_2(\pi|\mathbf{X})$ and conditional prior $f_2(\pi|\mu = 0.25)$]

[Figure 3 about here]

[Caption for Figure 3: Finite-sample posteriors $f_3(\pi|\mathbf{X})$ and conditional prior $f_3(\pi|\mu = 0.25)$]

From figures 4 and 5, it is clear that as the probability (λ) of error symmetry increases, the posterior distribution puts more and more mass on values less than 0.25. This occurs because under symmetry the restriction $\pi \leq \mu$ must hold. In the limit as $n \to \infty$ (see figure 6), the posterior becomes discontinuous at $\mu_0 = 0.25$ and values of π less than μ_0 are much more likely than values greater than μ_0. Again, this occurs because the conditional prior of π given μ has restricted support under symmetry of the errors and unrestricted support when the errors are asymmetric.

[Figure 4 about here]

[Caption for Figure 4: Finite-sample posteriors $f_4(\pi|\mathbf{X})$ for different probabilities (λ) of error symmetry; $n = 20$.]

[Figure 5 about here]

[Caption for Figure 5: Finite-sample posteriors $f_4(\pi|\mathbf{X})$ for different probabilities (λ) of error symmetry; $n = 100$.]

[Figure 6 about here]

[Caption for Figure 6: Conditional priors $f_4(\pi|\mu = 0.25)$ for different probabilities (λ) of error symmetry.]

The classical bounding results do not reveal anything about the location of the parameter within the identified set. In contrast, our Bayesian analysis is informative in that, under a range of prior distributions, the posterior of π has a unique mode. Moreover, the analysis provides insights into how different prior beliefs about the probability of a misclassification error affect the posterior distribution of π. As such, the Bayesian approach facilitates sensitivity analysis with respect to varying prior beliefs. Finally, as expected, the use of stronger information about misclassification rates will lead to more concentrated posterior distributions and tighter highest posterior density intervals.

5. Conclusion

In this paper we have derived a number of exact, finite-sample posterior distributions for the mean π of a misclassified binary variable. Despite the fact that the parameter π is not identified (unless the probabilities of misclassification errors are known), its posterior distribution provides non-trivial information even when very weak priors are specified. A Bayesian approach allows additional identification or sensitivity analysis in these set-identified models. Classical analyses often consider how the set of feasible values for the parameter of interest changes when certain model assumptions are either imposed or relaxed. In contrast, a Bayesian analysis allows researchers to impose or relax assumptions in a probabilistic and hence, more continuous manner. Such an analysis adds to our understanding of the mapping between assumptions and identification.

Acknowledgments

We would like to thank Robin Carter, Charles Manski, Tony Lancaster and participants at the SBIES and MEG conferences for valuable comments. We are solely responsible for any remaining errors.

References