MIS Quarterly

METHODS ARTICLE

CORRECTING MEASUREMENT ERROR IN REGRESSION MODELS WITH VARIABLES CONSTRUCTED FROM AGGREGATED OUTPUT OF DATA MINING MODELS[1]

Mengke Qiao
Culverhouse College of Business, University of Alabama,
Tuscaloosa, AL, U.S.A.{mqiao@ua.edu}

Ke-Wei Huang
School of Computing, National University of Singapore,
SINGAPORE {huangkw@comp.nus.edu.sg}

The burgeoning interest in data mining has catalyzed a proliferation of innovative techniques in extracting useful information from unstructured data sources, such as text and images in social sciences. One typical research design involves a two-stage process. In the first stage, researchers apply the classification algorithm to predict an individual-level categorical variable. In the second stage, researchers aggregate the predicted values to construct a group-level variable for further regression analysis. For example, text classification has been applied to classify whether a review is positive or negative. The predicted review sentiment is aggregated at the product level as a focal independent variable in a regression model to examine the impact of the average review sentiment on product sales. Since the first-stage classification will inevitably have errors, the aggregated variable may suffer from a measurement error in the regression analysis. Our study attempts to systematically investigate the theoretical properties of the estimation bias and introduce solutions rooted in theory to mitigate the issue of measurement error. We propose one exact solution and two approximated solutions based on the central limit theorem (CLT) and the law of large numbers (LLN), respectively. Our theoretical analysis and experimentation confirm that the consistency of regression estimators can be recovered across all examined scenarios and the approximated solutions offer a significantly reduced computational complexity compared to the exact solution. We also provide heuristic guidelines to choose one of three solutions.

Keywords: Data mining, aggregate-level regression model, measurement error, central limit theorem, law of large numbers

Introduction

The swift advancement of data mining has led social scientists to use classification algorithms to create new variables in regression analysis. Increasingly, researchers have been proposing various novel uses of classification algorithms to construct new variables from unstructured data. For example, text classification has been applied to classify whether a review is subjective or objective (Ghose & Ipeirotis, 2011; Ghose et al., 2012). The derived review subjectivity is employed as a focal independent variable in the regression to examine its impact on review helpfulness or product sales. Image classification is also gaining popularity. For example, image classification has been applied to classify whether the worker is male or female (Chan & Wang, 2018). The estimated gender is utilized as a focal independent variable in the regression to examine its impact on hiring outcomes.

[1] Sumit Sarkar was the accepting senior editor for this paper. Nachiketa Sahoo served as the associate editor. The transparency material for this paper is available at https://osf.io/72xr9/?view_only=4deca4c992794de6a4eeaee3671205d2

DOI:10.25300/MISQ/2024/18026

There are typically two stages in these “hybrid studies,” which are defined as using a first-stage classification algorithm to construct the focal independent or dependent variable for second-stage regression analysis. Throughout this paper, we use the term *predictive stage* to refer to the first stage and the *econometric stage* for the second stage. More specifically, in the predictive stage, researchers apply a classification algorithm on a small set of observations with labels (labeled set) to build a classifier. This classifier is applied to the unlabeled dataset to construct a new categorical variable, such as the (predicted) sentiment of a document. In the econometric stage, this newly constructed categorical variable can be directly utilized as the focal independent or dependent variable in a regression. For example, Liu et al. (2020) classified videos based on whether they encode a high or low degree of medical information. The classified output was directly used to predict the collective engagement with the video. In this paper, this type of hybrid study is defined as the “individual-level hybrid study”, where the predictive and econometric stages are analyzed at the same unit of analysis. Yang et al. (2018) and Qiao and Huang (2021) have proposed solutions for this case. However, it is not rare in the IS literature that the unit of analysis in the predictive stage and econometric stage could be different. It is quite common to use the mean or sum of the outcome variable from the predictive stage in the econometric model of the second stage. For example, Huang et al. (2019) examined the impact of the social capital of a member on the member’s provision of informational and emotional support in online support communities in the healthcare domain. In the predictive stage, Huang et al. (2019) classified each online posting message as “informational support” or “emotional support.” In the econometric stage, the regression analysis is conducted at the member level, not at the message level. As a result, the total number of “informational support” or “emotional support” messages posted by one member was constructed to measure the support provision of one member and used as the dependent variable in the econometric stage. In other words, the sum of the output variable of the first stage (not the output variable itself) is used as the dependent variable in the regression model. This type of hybrid study is defined as the “aggregate-level hybrid study,” where the econometric stage is conducted at the aggregate level, while the predictive stage is at the individual level.

One distinct feature of hybrid studies is that the constructed variable includes classification error because the output variable from the predictive stage cannot be classified perfectly. In the econometric stage, it has been well-documented in the econometrics literature that the measurement error of the independent or dependent variable may affect the estimation results. Yang et al. (2018) is the first paper investigating the measurement error issue in individual-level hybrid studies in the IS field. They adopted two simulation-based methods to correct the estimation bias, SIMEX, and MC-SIMEX, which can be parameterized using performance metrics from data mining models, such as error variance or the confusion matrix.

However, this problem becomes more complicated for aggregate-level hybrid studies and the solutions from the existing literature are not directly applicable due to the following reasons. In aggregate-level studies, we do not know the true values of the aggregated output variable from the prediction stage. In contrast, in individual hybrid studies, we know the true values of the individual output variable in the labeled dataset for classifier training and validation. As a result, in individual-level studies, we can quantify the measurement error model more precisely by the labeled dataset whereas, in aggregate-level studies, we can only estimate the distribution of measurement error probabilistically. For example, suppose researchers want to examine the impact of the average review sentiment on product sales. Unless the reviews of one product are all annotated, we cannot know the true values of the average review sentiment at the product level. Because in almost all hybrid studies, researchers randomly selected a small subset of reviews to be annotated, it is unlikely that we will know the labels of all reviews of one product. As a result, we need to derive a new solution for the aggregated hybrid study.

The aggregated level problem is also worthy of studying because the solutions are different when the aggregation function is mean or sum, which is equally common in the IS literature. Huang et al. (2019) is an example of the sum function. As an example of mean function, Deng et al. (2018) examined the impact of stock returns on microblog sentiment. In the predictive stage, they classified each microblog message as a “positive” or “nonpositive” message. Next, they constructed the “positive sentiment score” by the percentage of positive messages among all the messages, which was employed as the dependent variable in the regression. In other words, the mean of the first stage’s output variable is used as the dependent variable.[2] For ease of exposition, our paper uses the term *proxy variable* to refer to the predicted output label of the classification algorithm. We use the term *true variable* to refer to the true label predicted by the proxy variable. This study will investigate four cases. The aggregation function can be mean or sum and

[2] When the label of the message has three classes (“positive”, “negative”, or “neutral”), “positive sentiment score” is equivalent to the mean of the predicted message labels by combining “negative” and “neutral” messages as “non-positive” messages. In this paper, we represent “percentage or count of labels” by “mean or sum of labels” for notation simplification.

the aggregated variable can be used as the dependent or focal independent variable in a regression model. We will analyze the estimation bias and derive the theoretical solutions for all four cases.

The main contribution of our paper is to propose a new estimation procedure for aggregate-level hybrid studies, which is gaining popularity across social science disciplines. In the proposed solution, there are two important steps. First, researchers are recommended to train classifiers to meet several unique criteria, which are required by the "assumptions" in the econometric stage. Roughly speaking, researchers should try to minimize the correlation between the classification error and any other variables, such as the dependent variable in regression analysis. After deciding on the classifier, researchers can apply our theory-grounded solutions for aggregate-level hybrid studies to estimate the regression model. The main innovation of our method is we utilize the individual-level measurement error model and confusion matrix of each aggregated group to derive the aggregate-level measurement error model. Based on this model, we propose three solutions, including an exact solution and two approximated solutions based on the central limit theorem (CLT) and the law of large numbers (LLN). Our theoretical analysis shows that the consistency of regression estimators can be recovered in all cases studied in this paper and the time complexity of the approximated solutions is much better than the exact solution. Our study also contributes to the literature by providing theoretical analysis that quantifies the estimation bias if the researchers simply utilize the aggregated proxy variable without any correction (called the naïve approach in this paper) in the econometric stage. There are several counterintuitive findings. For example, when the mean of the proxy variable is used as the focal independent variable and the aggregation sample is large enough, the coefficient could be overestimated, which is different from the attenuation bias due to the classical measurement error in the traditional statistics literature. Finally, this paper also contributes to providing conditions under which researchers can ignore the measurement error.

The remaining paper is organized as follows. First, we provide a literature review. Next, we report a theoretical analysis of the estimation bias. Then, we report the main theoretical solutions for four cases and evaluate the proposed solutions through simulation studies and real-world applications. Our results show that our method can indeed correct the estimation inconsistency. Finally, we conclude this paper.

Literature Review

Aggregate-Level Hybrid Study Applications

Recently, applying supervised machine learning methods to construct variables from unstructured data has gained popularity, which has facilitated the popularity of hybrid studies in the information systems discipline (Chen et al., 2012). Abundant examples of aggregated-level hybrid studies exist in the IS literature. For example, the text label of each online review can be aggregated at the product level on e-commerce websites (Ghose & Ipeirotis, 2011; Wu et al., 2019) or at the seller level in online service marketplaces (Moreno & Terwiesch, 2014). Besides the text label, on Airbnb, the image label of each room is also aggregated at the property level (Zhang et al., 2016). Similarly, extensive examples exist in the online community literature. For example, the labels of individual online postings have been aggregated at the solver level in user support (Q&A) forums (Jabr et al., 2014), at the stock level and at the discussion thread level in online communities for investment (Gu et al., 2007, 2014), at the IT venture level in online blog platform (Aggarwal & Singh, 2013), at the brand level in the social media platform (Luo et al., 2013), at the member level in healthcare virtual support communities (Huang et al., 2019), and at the topic level in the enterprise blogosphere (Singh et al., 2014). In summary, the constructed variable from the classification algorithm can be aggregated at various higher levels depending on research contexts.

Table 1. Summary of Four Cases of Aggregate-Level Hybrid Studies

	Aggregation form of the proxy variable	Example	Aggregated variable	Model
Case1	mean as IV	Moreno & Terwiesch (2014)	Mean of comment labels	Multinomial logit model
Case2	sum as IV	Gu et al. (2007)	Sum of post labels	Multinomial logit model
Case3	mean as DV	Deng et al. (2018)	Mean of message labels	Vector autoregression
Case4	sum as DV	Jabr et al. (2014)	Sum of post labels	Negative binomial model

There are typically four cases of aggregate-level hybrid studies, as summarized in Table 1. In the first case, researchers utilize the mean of the individual-level constructed variable as the focal independent variable in the econometric stage. For example, Moreno and Terwiesch (2014) examined the impact of the reputation score in online service marketplaces on buyers' and sellers' behavior. In the predictive stage, they classified each comment received by the seller as "positive" or "negative." Next, they defined a new variable "reputation score" of each seller by the percentage of "positive" comments among all comments. In the econometric stage, the "reputation score" was used as the focal independent variable in the main regression model. In the second case, researchers utilize the sum of the individual-level variable as the focal independent variable in the econometric model. For example, Gu et al. (2007) examined the impact of the number of quality postings of one stock on the user's choice of online communities. In the predictive stage, they classified each posting as "noise", "neutral" or "signal." In the econometric stage, the count of "signal" posts for one stock was used as the focal independent variable in the econometric model. Because this paper analyzes the network effect of online communities, it makes more sense to use the sum rather than the mean as the aggregation function. The main implication is that while it seems more intuitive to use mean as the aggregation function, there exist cases where researchers have to use sum as the aggregation function for theoretical reasons. An example of the third case is Deng et al. (2018), who examined the impact of stock returns on microblog sentiment. In the predictive stage, they classified each microblog message as "positive" or "non-positive." Next, they defined the daily "positive sentiment score" by the ratio of the number of positive messages to the total number of messages on that day. In the econometric stage, the "positive sentiment score" was used as the dependent variable in the regression model. An example of the fourth case is Jabr et al. (2014), who examined how the contribution of solvers is affected by the recognition mechanism in user support (Q&A) forums. In the predictive stage, they classified each post as a solution post or not. Next, they defined a new variable "contribution level" of each solver by the number of solution posts. In the econometric stage, the "contribution level" was the key dependent variable in the regression. In this paper, all four cases will be analyzed.

Measurement Error of Variables in Aggregate-level Hybrid Studies

Yang et al. (2018) and Qiao and Huang (2021) have investigated measurement error in individual-level hybrid studies in the IS field. To correct the estimation inconsistency, in their pioneering paper, Yang et al. (2018) utilized two simulation-based methods, SIMEX and MC-SIMEX, which were applied to continuous variables with additive measurement error and discrete variables with misclassification (Cook & Stefanski, 1994; Küchenhoff et al., 2006). Qiao and Huang (2021) proposed theoretical solutions to correct the misclassification bias for generalized linear models. However, the proposed methods in both papers cannot be applied in aggregate-level hybrid studies since both solutions require the aggregate-level measurement error model, which cannot be directly quantified from the labeled dataset.

The issue of measurement error has been widely studied by econometricians (Greene, 2012) and biostatisticians (Buonaccorsi, 2010). Abundant methods have been proposed to correct the estimation inconsistency due to measurement error. However, most of the existing literature has mainly focused on solutions involving researchers diagnosing the measurement error model of the mismeasured variable directly. Buonaccorsi (2010) and Carroll et al. (2006) offer detailed analyses. Little literature has analyzed the aggregated measurement error issue with a unique structure where the aggregation step can cancel out or aggravate the individual-level measurement error. Fuller (1987) examined the measurement error issue in the independent variable when beta coefficients are different for separate groups in the data. However, this combination of grouping and measurement error is different from our setting.

Theoretical Analysis of Estimation Bias Due to Aggregated Measurement Error ■

In this section, we first explain and define the required notations of our model. Next, we characterize the mean and variance of the measurement error. Then, we theoretically analyze estimation bias in the linear regression when the measurement error is ignored, as in the naïve approach widely adopted in the existing literature. Finally, we examine the impacts of measurement error in different regression models by simulation.

Notations and Definitions

In this paper, we consider only one focal variable $\bar{X}_i$, which is the mean of the *true variable* (X_i^j) for the aggregated group *i*. M_i is the number of observations for aggregating to derive $\bar{X}_i$. X_i^j is the true label of row *j* in aggregated group *i*. For example, if researchers wanted to examine the impact of the average review sentiment at the product level on product sales, the average review sentiment ($\bar{X}_i$) would be calculated based on the sentiment labels (X_i^j) of all the reviews received by product *i*. We define the reviews received by product *i* as the aggregated group *i*. M_i is the total number of reviews received by product *i* (group *i*). X_i^j is the true sentiment label of review *j* in product group *i.* The value of X_i^j is unknown and is predicted by the output in the predictive stage. W_i^j is named the *proxy variable* of X_i^j, which is constructed by predicted values from the classifier in the predictive stage. In hybrid studies, researchers do not know the values of X_i^j (except in a small, labeled dataset for training a classifier), while W_i^j is known for all the records. In almost all existing hybrid studies, researchers estimated the regression model by replacing $\bar{X}_i$ with $\bar{W}_i$ given the implicit assumption that $\bar{X}_i$=$\bar{W}_i$, which does not hold in most scenarios. In this study, we define a classifier as unbiased when $\mathrm{E}(\bar{X}_i)$ = $\mathrm{E}(\bar{W}_i)$ for all *i*. Based on this view, in the existing hybrid studies, researchers imposed two strong assumptions: (1) all classifiers are unbiased and (2) M_i is large enough. To the best of our knowledge, none of the existing hybrid studies have discussed whether $\mathrm{E}(\bar{X}_i)$=$\mathrm{E}(\bar{W}_i)$ for all *i* is a valid assumption in their empirical studies. Similarly, when the aggregation function is the sum, we define the sum of the true variable and proxy variable by S_{X_i} and S_{W_i}.

To analyze the aggregated measurement error, we need to utilize the information from the aggregate-level confusion matrix. We illustrate the confusion matrix of *one* aggregated group *i* in Table 2. Because this is only for one group of data, the total number of cases is M_i, which can be decomposed into four important numbers: true positive cases (TP), true negative cases (TN), false positive cases (FP), and false negative cases (FN). We define h as the number of true positive observations. We can express all terms in Table 2 using h and three other notations M_i, $\bar{W}_i$, and $\bar{X}_i$. With these terms, we can also express all probability terms by these four variables. For example, we define two probabilities, $\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} = h/M_i\bar{X}_i$ and $\widehat{\mathrm{Pr}}_i^{\mathrm{FP}} = (M_i\bar{W}_i - h)/(M_i - M_i\bar{X}_i)$. These two values are also the two axes in a typical ROC figure.

Expected Value and Variance of Aggregated Measurement Error

Given the notations in Table 2, it is straightforward to derive the measurement error between $\bar{W}_i$ and $\bar{X}_i$ because $\bar{W}_i$ can be rewritten as a function of $\bar{X}_i$ as follows:

$$\bar{W}_i = \bar{X}_i \times \widehat{\mathrm{Pr}}_i^{\mathrm{TP}} + (1 - \bar{X}_i) \times \widehat{\mathrm{Pr}}_i^{\mathrm{FP}}.$$

Subtracting both sides by $\bar{X}_i$, we can derive an expression of the measurement error as:

$$\bar{e}_i = \bar{W}_i - \bar{X}_i = \bar{X}_i \times \left(\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + (1 - \bar{X}_i) \times \widehat{\mathrm{Pr}}_i^{\mathrm{FP}}.$$

We define the expected values of $\widehat{\mathrm{Pr}}_i^{\mathrm{TP}}$ and $\widehat{\mathrm{Pr}}_i^{\mathrm{FP}}$ by $\mathrm{E}(\widehat{\mathrm{Pr}}_i^{\mathrm{TP}})$= $\overline{\mathrm{Pr}}_i^{\mathrm{TP}}$ and $\mathrm{E}(\widehat{\mathrm{Pr}}_i^{\mathrm{FP}})$= $\overline{\mathrm{Pr}}_i^{\mathrm{FP}}$**.** Then the expectation and variance of the measurement error can be derived as:

$$\mathrm{E}(\bar{e}_i) = \mathrm{E}(\bar{X}_i) \times \left(\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + \left(1 - \mathrm{E}(\bar{X}_i)\right) \times \overline{\mathrm{Pr}}_i^{\mathrm{FP}},$$

$$\mathrm{Var}(\bar{e}_i) = \frac{\mathrm{E}(\bar{X}_i)\overline{\mathrm{Pr}}_i^{\mathrm{TP}}\left(1 - \overline{\mathrm{Pr}}_i^{\mathrm{TP}}\right) + \left(1 - \mathrm{E}(\bar{X}_i)\right)\overline{\mathrm{Pr}}_i^{\mathrm{FP}}\left(1 - \overline{\mathrm{Pr}}_i^{\mathrm{FP}}\right)}{M_i} + (\overline{\mathrm{Pr}}_i^{\mathrm{FN}} + \overline{\mathrm{Pr}}_i^{\mathrm{FP}})^2\mathrm{Var}(\bar{X}_i).$$

The proof is shown in Appendix C. Following a similar logic, we can derive the expectation and variance of the error of the sum S_{W_i} relative to S_{X_i} conditional on M_i as follows

$$\mathrm{E}\left(S_{e_i}\right) = M_i\mathrm{E}(\bar{e}_i), \mathrm{Var}(S_{e_i}) = {M_i}^2\mathrm{Var}(\bar{e}_i).$$

In summary, the expectation and variance of measurement error in both cases can be fully characterized by the two proportions of TP and FP because $\mathrm{E}(\bar{X}_i)$ results from the nature of the dataset and is invariant to the classifier performance. Another important implication is that the expectation of the error is zero only when $\mathrm{E}(\bar{X}_i) \times \left(\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + \left(1 - \mathrm{E}(\bar{X}_i)\right) \times \overline{\mathrm{Pr}}_i^{\mathrm{FP}} = 0$. If the proportion of 0 is 50%, this expression simplifies to $\left(\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + \overline{\mathrm{Pr}}_i^{\mathrm{FP}} = 0$. In other words, only when the classifier satisfies this condition for all groups will the expectation of the error be zero, and this is unlikely to happen in practice. This is one important empirical condition that has been overlooked by researchers in the literature.

Estimation Bias Due to Aggregated Measurement Error

Review of Estimation Bias Due to Measurement Error in the Traditional Literature

We consider a simple linear regression with only one regressor: $Y_i = X_i\beta + \varepsilon_i$. ε_i is the error term and is independent of X_i. We assume that X_i is unobserved, and we observe W_i instead. Following the standard model setup in the econometrics literature, $W_i = X_i + e_i$. e_i is the classical measurement error, independent of both X_i and ε_i. Then the regression model becomes $Y_i = W_i\beta + \varepsilon_i - e_i\beta$. Let N denote the sample size of the dataset. If we ignore e_i in the regression analysis, the estimated β will have attenuation bias (Greene, 2012),

$$\lim_{N\to\infty} \hat{\beta} = \frac{\mathrm{Cov}(W_i\beta, W_i)}{\mathrm{Var}(W_i)} + \frac{\mathrm{Cov}(-e_i\beta, W_i)}{\mathrm{Var}(W_i)} + \frac{\mathrm{Cov}(\varepsilon_i, W_i)}{\mathrm{Var}(W_i)} = \beta + \frac{\mathrm{Cov}(-e_i, W_i)}{\mathrm{Var}(W_i)}\beta = \beta - \frac{\sigma_e^2}{\sigma_X^2 + \sigma_e^2}\beta, \tag{1}$$

where σ_X^2 and σ_e^2 are the variance of X_i and e_i, and $\mathrm{Cov}(\varepsilon_i, W_i) = 0$.

Table 2. Confusion Matrix for Predicted Variable

	$\boldsymbol{X_i^j = 1}$	$\boldsymbol{X_i^j = 0}$	**Sum**
$\boldsymbol{W_i^j = 1}$	TP (h)	FP ($M_i\overline{W}_i - h$)	($M_i\overline{W}_i$)
$\boldsymbol{W_i^j = 0}$	FN ($M_i\bar{X}_i - h$)	TN (($M_i - M_i\overline{W}_i$) − ($M_i\bar{X}_i - h$))	($M_i - M_i\overline{W}_i$)
Sum	($M_i\bar{X}_i$)	($M_i - M_i\bar{X}_i$)	M_i

Next, we assume Y_i has the classical measurement error, $y_i = Y_i + e_i$. Then the regression model becomes $y_i = X_i\beta + \varepsilon_i + e_i$. In this case, if we ignore e_i in the regression analysis, there is no bias in the estimated coefficient (Greene, 2012),

$$\lim_{N\to\infty} \hat{\beta} = \frac{\mathrm{Cov}(X_i\beta, X_i)}{\mathrm{Var}(X_i)} + \frac{\mathrm{Cov}(e_i, X_i)}{\mathrm{Var}(X_i)} + \frac{\mathrm{Cov}(\varepsilon_i, X_i)}{\mathrm{Var}(X_i)} = \beta + \frac{\mathrm{Cov}(e_i, X_i)}{\mathrm{Var}(X_i)} = \beta. \quad (2)$$

In the following sections, our focus is on analyzing the estimation bias in simple linear regression when the measurement error of the proxy variable is aggregated.

Mean of Proxy Variable as the Focal Independent Variable

When the mean of the proxy variable is the focal independent variable, the measurement error is averaged, $\bar{e}_i = \overline{W}_i - \bar{X}_i$. The regression model becomes $Y_i = \overline{W}_i\beta + \varepsilon_i - \bar{e}_i\beta$. Different from traditional literature, we further decompose $\bar{e}_i$ as:

$$\bar{e}_i = \mathrm{E}(\bar{e}_i) + \mu_i,$$

where $\mathrm{E}(\bar{e}_i) = -\left(\frac{1}{\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - \overline{\mathrm{Pr}}_i^{\mathrm{FP}}} - 1\right)\mathrm{E}(\overline{W}_i) + \frac{\overline{\mathrm{Pr}}_i^{\mathrm{FP}}}{\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - \overline{\mathrm{Pr}}_i^{\mathrm{FP}}}$, the expectation of μ_i is 0, and the variance of μ_i is the variance term in the Expected Value and Variance of Aggregated Measurement Error section. The proof is shown in Appendix C.

Next, we derive the formula of estimated β by replacing W_i by $\overline{W}_i$ and e_i by $\bar{e}_i$ in Equation (1),

$$\begin{aligned}\lim_{N\to\infty} \hat{\beta} &= \beta + \frac{\mathrm{Cov}(-\bar{e}_i, \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)}\beta \\ &= \beta + \frac{\mathrm{Cov}(-\mathrm{E}(\bar{e}_i), \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)}\beta + \frac{\mathrm{Cov}(-\mu_i, \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)}\beta.\end{aligned}$$

We assume that $\overline{\mathrm{Pr}}_i^{\mathrm{TP}}$ and $\overline{\mathrm{Pr}}_i^{\mathrm{FP}}$ are the same across all the aggregated groups. This assumption is also imposed in the subsequent part. Then we can remove subscript *i* for $\mathrm{E}(\bar{e}_i)$ and obtain:

$$\begin{aligned}\lim_{N\to\infty} \hat{\beta} = \beta + \left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} - 1\right)\frac{\mathrm{Cov}(\mathrm{E}(\overline{W}_i), \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)}\beta \\ + \frac{\mathrm{Cov}(-\mu_i, \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)}\beta,\end{aligned}$$

where *N* is the sample size of the full dataset for regression. When M_i is large enough, $\overline{W}_i$ and μ_i converge to $\mathrm{E}(\overline{W}_i)$ and 0 by the law of large numbers. Therefore, $\frac{\mathrm{Cov}(\mathrm{E}(\overline{W}_i), \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)} = 1$ and $\frac{\mathrm{Cov}(-\mu_i, \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)} = 0$. The coefficient is overestimated as follows,

$$\lim_{N\to\infty, M_i\to\infty} \hat{\beta} = \beta + \beta\left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} - 1\right) = \frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}}\beta.$$

In summary, when the mean of the proxy variable is used as the focal independent variable in simple linear regression, the coefficient varies with M_i and is overestimated by $\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}}$ when *N* and M_i are large enough. This result is different from the attenuation bias due to the classical measurement error of the independent variable in the existing measurement error literature in the Review of Estimation Bias Due to Measurement Error in the Traditional Literature section. The reason is that the classical measurement error ($-e_i$) is negatively correlated with W_i, while the error of the mean of the proxy variable ($-\bar{e}_i$) is positively correlated with $\overline{W}_i$ when M_i is large enough.

Mean of Proxy Variable as the Dependent Variable

When the mean of the proxy variable is utilized as DV, $\bar{e}_i = \bar{y}_i - \bar{Y}_i$. The model becomes $\bar{y}_i = X_i\beta + \varepsilon_i + \bar{e}_i$. Next, the last term is:

$$\bar{e}_i = \mathrm{E}(\bar{e}_i) + \mu_i,$$

where $\mathrm{E}(\bar{e}_i) = X_i\beta \times (\overline{\mathrm{Pr}}^{\mathrm{TP}} - 1) + (1 - X_i\beta) \times \overline{\mathrm{Pr}}^{\mathrm{FP}}$ conditional on X_i.

Finally, we derive the formula of estimated β by replacing e_i with $\bar{e}_i$ in Equation (2):

$$\begin{aligned}\lim_{N\to\infty} \hat{\beta} = \beta + \frac{\mathrm{Cov}(\bar{e}_i, X_i)}{\mathrm{Var}(X_i)} = \beta + \frac{\mathrm{Cov}(\mathrm{E}(\bar{e}_i), X_i)}{\mathrm{Var}(X_i)} + \frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)} \\ = (\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}})\beta.\end{aligned}$$

The third equality results from the fact that $\frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)} = 0$. The detailed proof for this analysis is shown in Appendix C. Therefore, when the mean of the proxy variable is used as the dependent variable in simple linear regression, the coefficient does not vary with M_i and is underestimated by $\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}$. This result is different from the no-bias results in the traditional statistics literature. The reason is that the literature assumes that the measurement error is not correlated with the independent variable, while in our case, the aggregated measurement error is negatively correlated with the independent variable. The estimation bias results for the sum of proxy variable cases have a slight difference from the mean cases, which are shown in Appendix C due to the page limit.

Estimation Bias Simulation Results

We also used simulation to characterize how the estimation bias varies with the sample size of the aggregated group and the nature of measurement error in different regression models (including OLS, logit model, and beta-binomial model). Due to the page limit, we omitted the results, which are available upon request. The results are consistent with our theoretical analysis. Generally, the estimation bias decreases when the classification error is smaller. Moreover, when the mean or sum of the proxy variable is used as the focal independent variable, the estimation bias increases from “underestimated” to “overestimated” as the sample size of the aggregated group increases. When the mean or sum of the proxy variable is used as the dependent variable, the coefficient is underestimated and the estimation bias does not change as the sample size of the aggregated group increases.

Theoretical Solutions

Theoretical Solution Framework

Our method is built on the traditional probabilistic approach by using the maximum likelihood estimation (MLE) method for correcting the measurement error issue in Carroll et al. (2006). We first introduce a general solution based on this approach.

True Model in the Econometric Stage

We denote the econometric stage’s true model by $\mathrm{P}(Y|\bar{X}, Z; \theta)$, which is the probability function of Y conditional on $\bar{X}$ and Z. The most common example is logistic regression. To illustrate the flexibility of this approach, we will analyze several cases of this function, including the widely used generalized linear model (GLM). In this specification, Y can be interpreted as the dependent variable, $\bar{X}$ can be interpreted as the focal independent variable, and Z is the vector of the control variables. θ is a vector of regression parameters for estimation. Testing whether θ is zero is the objective in hybrid studies. However, in hybrid studies, researchers do not know the values of $\bar{X}$, while $\bar{W}$ is known for all the records. Therefore, in almost all existing literature, researchers simply replace $\bar{X}$ with $\bar{W}$ in $\mathrm{P}(Y|\bar{X}, Z; \theta)$ and proceed with the standard estimation procedure using OLS or MLE. Our theoretical and simulation results show that this approach suffers from the measurement error problem, leading to the inconsistent estimation of θ.

A General Solution by MLE in the Econometric Stage

The correct objective function for MLE should not be replacing $\bar{X}$ with $\bar{W}$ in $\mathrm{P}(Y|\bar{X}, Z; \theta)$ under the implicit assumption that $\bar{X} = \bar{W}$. Instead, $\bar{W}$ and other observable variables should be used to estimate the probability distribution of $\bar{X}$. The correct objective function is a weighted sum of candidates of true values of $\bar{X}$. Formally, the weight is denoted by $\mathrm{P}(\bar{X}|\bar{W}, Z)$, which is the probability function of $\bar{X}$, conditional on $\bar{W}$ and Z. We define this term as the *measurement error* model. Empirically, this function needs to be estimated, and this study proposes three methods to estimate this probability function. To explain this issue using the simplest example, we consider a dataset with only two textual reviews for each product and use the mean of these two binary sentiment values as the focal independent variable. When we observe $\bar{W} = 1$, we need to estimate the probability that the true mean $\bar{X}$ could be 0, 0.5, or 1. Most researchers simply “assume” $\bar{X} = \bar{W} = 1$ with 100% probability and proceed with MLE for logistic regression. However, to correct the measurement error of $\bar{X}$, the correct approach is to estimate the conditional probability that $\bar{X}$ could be 0, 0.5, or 1 and use the following function for MLE:

$$\mathrm{P}(Y|\bar{W}, Z) = \sum_{\bar{X}} \mathrm{P}(Y|\bar{X}, Z, \bar{W}) \times \mathrm{P}\,(\bar{X}|\bar{W}, Z). \tag{3}$$

Intuitively, this equation means that we compute $\mathrm{P}(Y|\bar{W}, Z)$ by taking the expectation of the conditional probability function $\mathrm{P}(Y|\bar{X}, Z, \bar{W})$ over all possible values of $\bar{X}$. In the following discussion, the focus is on how to estimate the measurement error function $\mathrm{P}(\bar{X}|\bar{W}, Z)$.

Case 1: Mean of Proxy Variable as the Focal Independent Variable

We first analyze the case in which the mean of the proxy variable is the focal independent variable in the generalized linear model in the econometric stage. This may be the most widely analyzed econometric model among the four cases in the existing literature. The dependent variable in the econometric stage is denoted by Y_i, which is an *N*-by-1 vector. The unobservable independent variable is denoted by $\bar{X}_i$, which is the mean of the true variable for the aggregated group *i*. $\bar{X}_i$ is also an *N*-by-1 vector. M_i is the number of observations for aggregating to derive $\bar{X}_i$. X_i^j is the true label of row *j* in the predictive stage and the mean of X_i^j in group i is $\bar{X}_i$. For example, Moreno and Terwiesch (2014) examined the impact of the reputation score in online service marketplaces on buyers' and sellers' behavior. The reputation score ($\bar{X}_i$) was defined as the average sentiment of the review comments received by seller *i* in previous projects. The text classification was applied at the comment level to predict the sentiment label. Next, the authors averaged the comment-level labels to construct the seller-level reputation score. M_i was the total number of comments received by seller i because different sellers may receive different numbers of comments. The unit of econometric analysis is at the seller level.

True Model in the Econometric Stage

Our proposed solution is applicable to most of the popular econometric models with a probability function that could be estimated by MLE. One widely used model is the generalized linear model (GLM), which is the model explained here. In this case, the ideal regression is specified as:

$$\begin{aligned} &\mathrm{P}(Y_i|\bar{X}_i, Z_i) = G(\bar{X}_i\beta + Z_i\gamma), \\ &\bar{X}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} X_i^j,\ i = 1, \dots, N, \end{aligned} \tag{4}$$

where $G()$ is the link function (such as logistic or probit function) and β is the focal regression coefficient for estimation, Z_i is an *N*-by-*K* matrix of control variables. γ is a *K*-by-1 vector of regression coefficients of control variables.

Measurement Error Model 1: Exact Solution by Binomial Distribution

This section explains how to estimate the measurement error model, which is the core of the solution. In hybrid studies, X_i^j is unobservable and only W_i^j is observable. Most hybrid studies estimate the following regression model by replacing $\bar{X}_i$ with $\bar{W}_i$:

$$\mathrm{P}(Y_i|\bar{X}_i = \bar{W}_i, Z_i) = G(\bar{W}_i\hat{\beta} + Z_i\hat{\gamma}).$$

$\hat{\beta}$ is called the naïve estimator throughout this paper. However, the coefficients estimated from this model are not consistent because the probability function does not correctly account for the relationship between $\bar{X}_i$ and $\bar{W}_i$. The correct objective function for MLE is Equation (3). We make the following two assumptions to further simplify Equation (3).

Assumption 1: *$\bar{W}_i$ provides no additional information about Y_i conditional on $\bar{X}_i$ and Z_i.* In other words, we assume that $\mathrm{P}(Y_i|\bar{X}_i, Z_i, \bar{W}_i) = \mathrm{P}(Y_i|\bar{X}_i, Z_i)$.

Assumption 2: *Define $e_i^g = X_i^g - \mathrm{E}(X_i^g|W_i^g, Z_i, \bar{W}_i)$. We assume that e_i^g is uncorrelated with X_i^h and W_i^h for all $h \neq g$.*

Assumption 1 is similar to the standard assumption imposed in the measurement error literature in econometrics. In the textbook version of the measurement error model, the measurement error term does not correlate with the dependent variable.[3] In Assumption 1, we assume that the mean of the proxy variable does not provide additional information for predicting *Y*. The validity of this assumption depends more on the choice of classifier than the nature of the data. Recall that researchers can choose classification algorithms and hyper-parameter tuning to affect the values of W. Consequently, there is no business interpretation of $(W - X)$ because its value can be manipulated by researchers. This enables researchers to tune the chosen classifier so that Assumptions 1 and 2 are valid. As a result, the proposed theoretical formula in the econometric stage can produce a consistent estimator of the regression coefficient.

In Assumption 2, e_i^g represents the residual term between actual X_i^g and its conditional mean. Assumption 2 implies that this residual term does not correlate with the true values and predicted values of all other observations. The main reason for imposing Assumption 2 is to simplify the proof of theorems. Without this assumption, to estimate the conditional probability distribution of X_i^g, we need to consider other observations' predicted labels, and it becomes intractable to include all the predicted labels for predicting X_i^g since different aggregated groups have different numbers of observations. In general, it is possible that Assumptions 1 and 2 are violated and researchers are advised to make sure that the first stage's classifier does not produce a classification error that correlates with all other variables.[4]

[3] Assumption 1 implies that the misclassification error does not provide additional information for the regression analysis. From the perspective of the econometrics stage, this assumption greatly simplifies the analysis; therefore, this assumption is widely used in existing measurement error literature. For example, Buonaccorsi (2010) imposed a similar assumption when analyzing the individual-level measurement error issue in a generalized linear model. From the perspective of the data mining stage, there exists almost no literature that analyzes the relationship between the classification error and the target variable. In other words, in the data mining literature, researchers rarely discuss whether the classification error term correlates with the true value of the target variable for prediction because the focus is usually the prediction accuracy.

[4] For the verification of our assumption, researchers can test our assumption by estimating one model where X_i^{g-1} and W_i^{g-1} ($g-1$ means the observation that is next to the g^{th} observation in the group) are the independent variables and e_i^g is the dependent variable using the validation

Theorem 1: *Let the labeled dataset be an i.i.d random sample drawn from the population. Given Assumptions 1 and 2, β in Equation (4) can be consistently estimated by applying MLE to:*

$$\mathrm{P}(Y_i|\bar{W}_i, Z_i) = \sum_{\bar{X}_i} \mathrm{P}(Y_i|\bar{X}_i, Z_i)\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i), \tag{5}$$

where the measurement error function can be decomposed as follows:

$$\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i) = \sum_{h=0}^{M_i\bar{X}_i} \mathrm{B}\big(h; M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{TP}}\big)\times \mathrm{B}\big(M_i\bar{X}_i - h; M_i - M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{FN}}\big). \tag{6}$$

$\mathrm{B}\big(h; M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{TP}}\big)$ and $\mathrm{B}\big(M_i\bar{X}_i - h; M_i - M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{FN}}\big)$ are two binomial probability mass functions explained in the following paragraphs. In summary, with Assumption 1, we simplify Equation (3) to Equation (5). With Assumption 2, we decompose $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ using Equation (6). Detailed proof of this theorem is provided in Appendix A.[5] [6]

To derive the conditional probability $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ at the aggregated level, we rely on the classification confusion matrix from the predictive stage in Table 2. The four important numbers (TP, TN, FP, and FN) correspond to the only four possible values of the individual-level measurement error model $\mathrm{P}(X_i^j|W_i^j, Z_i, \bar{W}_i)$ since both X_i^j and W_i^j are binary variables.[7]

Specifically, the probabilities of TP and FN observations at the individual level are $\mathrm{P}(X_i^j = 1|W_i^j = 1, Z_i, \bar{W}_i)$ and $\mathrm{P}(X_i^j = 1|W_i^j = 0, Z_i, \bar{W}_i)$, which are defined as $\mathrm{Pr}_i^{\mathrm{TP}}$ and $\mathrm{Pr}_i^{\mathrm{FN}}$. M_i is the sample size of this aggregated group and we know that there are $M_i\bar{W}_i$ predicted positive observations. If the number of actual positive observations is $M_i\bar{X}_i$ and h is the number of TP cases, then we must have $M_i\bar{X}_i - h$ observations of FN. Now we are ready to explain the intuition behind $\mathrm{B}\big(h; M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{TP}}\big)$ and $\mathrm{B}(M_i\bar{X}_i - h; M_i - M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{FN}})$ in Equation (6). Under the independence Assumption 2, the probability of "h TP observations out of $M_i\bar{W}_i$ predicted positive observations" can be modeled by a binomial distribution, $\mathrm{B}(h; M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{TP}})$ with the first parameter being the number of independent draws (total number of predicted positive observations) and the second parameter being the probability of the event (TP rate). Similarly, $\mathrm{B}\big(M_i\bar{X}_i - h; M_i - M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{FN}}\big)$ means the probability that there are $M_i\bar{X}_i - h$ FN observations out of $M_i - M_i\bar{W}_i$ predicted negative observations.

Next, we write the conditional probability function of obtaining this confusion matrix as:

$$\mathrm{P}(\bar{X}_i, h|\bar{W}_i, Z_i) = \mathrm{B}\big(h; M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{TP}}\big) \times \mathrm{B}\big(M_i\bar{X}_i - h; M_i - M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{FN}}\big).$$

Finally, we decompose $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ by all possible values of h (the number of true positive observations) and for each pair of $(\bar{X}_i, h)$, the probability is a multiplicative term of two binomial distribution probabilities. The numbers of predicted positive and negative observations are observable from the output of the first-stage prediction. The two terms $\mathrm{Pr}_i^{\mathrm{TP}}$ and $\mathrm{Pr}_i^{\mathrm{FN}}$ conditional on W_i^j, Z_i, and $\bar{W}_i$ can be estimated by applying logistic regression to the labeled set in the predictive stage with cross-validation. We also provide one example to explain how to calculate $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ (see Appendix A).

Therefore, given any classifier,[8] we can always estimate the individual-level measurement error model by using the validation dataset. Next, we use this individual-level result and the confusion matrix to estimate the aggregate-level measurement error model using Equation (6). Finally, we correct the estimation bias using Equation (5).[9] However, classifiers with lower accuracy produce the estimated coefficients with larger variances. We recommend that researchers select the classifier with the smallest variance.[10,11]

dataset. For verification purpose, researchers may need to manually label to obtain the values of X_i^{g-1}. If the coefficients of X_i^{g-1} and W_i^{g-1} are significant, then our assumption may be violated. The test for the DV case is similar.

[5] Our method can be applied to the GLM models and survival models by replacing $\mathrm{P}(Y_i|\bar{X}_i, Z_i)$ in Equation (5) with the corresponding probability function of a GLM or survival model.

[6] This solution can also be extended to the cases when multiple aggregated variables are the focal independent variables. Please refer to Appendix A for more details.

[7] This solution is applicable to the case when the outputs of the classifier have multiple classes (e.g., "positive", "negative", and "neutral"). In this case, X_i^j and W_i^j can be constructed as binary variables by combining "negative" and "neutral" as "non-positive" when the focal variable is the percentage of positive labels.

[8] The requirement is that the sum of the true positive rate and true negative rate of the classifier is larger than 1.

[9] Since our method involves two MLE steps where the first step estimates the measurement error model and the second step estimates the regression coefficient, we compute the variance based on the variance formula for two-step maximum likelihood estimation in Murphy and Topel (2002).

[10] Since this conclusion has been explained in Qiao and Huang (2021), we do not explain this point in detail.

[11] We suggest that it is better to have a larger labeled set and the dataset used in the second stage econometric model since the classifier training will be more accurate and the variance of the estimated coefficients will decrease as the sample sizes increase. While it is always beneficial from an estimation perspective to increase both sample sizes, this may not be practical if the total data set is fixed and predetermined. In this situation, it is feasible to increase the sample size of labeled set for enhancing the precision of the estimation.

The major drawback of the solution in Theorem 1 is the complexity issue arising when M_i becomes arbitrarily large. Fortunately, when M_i is large enough, we can apply the central limit theorem or law of large numbers to derive the simplified formula of $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$. In the next two sections, we analyze the large sample property of this problem.

Measurement Error Model 2: Approximated Solution by Normal Distribution

In this case, we maintain Assumptions 1 and 2. When M_i is large enough, the mean value of X_i^j may converge in distribution to a normal distribution by Lyapunov central limit theory (CLT) (Bentkus, 2005). In addition, our solution in Equation (6) relies on the binomial distribution, which is approximated by normal distribution asymptotically (Schader & Schmid, 1989).

Specifically, for predicted positive observations, the number of TP observations follows the binomial distribution $\mathrm{B}(\cdot\,; \mathrm{WP}, \mathrm{Pr}_i^{\mathrm{TP}})$, with WP equaling $M\bar{W}_i$. For predicted negative observations, the number of FN observations follows $\mathrm{B}(\cdot\,; \mathrm{WN}, \mathrm{Pr}_i^{\mathrm{FN}})$, with WN equaling $M - M\bar{W}_i$. When the sample size is large enough, two binomial distributions can be approximated by the normal distributions. Specifically, $\mathrm{B}(\cdot\,; \mathrm{WP}, \mathrm{Pr}_i^{\mathrm{TP}})$ is approximated by $\mathrm{N}(\cdot\,; \mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}}, \mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}} \times \mathrm{Pr}_i^{\mathrm{FP}})$ and $\mathrm{B}(\cdot\,; \mathrm{WN}, \mathrm{Pr}_i^{\mathrm{FN}})$ is approximated by $\mathrm{N}(\cdot\,; \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}}, \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}} \times \mathrm{Pr}_i^{\mathrm{TN}})$. Detailed proof of this part is provided in Appendix A. With a simple transformation, we can derive the normal density function of the mean $\bar{X}_i$. The mean and variance of this normal distribution are given by:

$$\mu_i = \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i) = \frac{1}{M_i}[\mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}} + \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}}],$$

$$\sigma_i^2 = \mathrm{Var}(\bar{X}_i|\bar{W}_i, Z_i) = \frac{1}{{M_i}^2}[\mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}} \times \mathrm{Pr}_i^{\mathrm{FP}} + \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}} \times \mathrm{Pr}_i^{\mathrm{TN}}],$$

where $\mathrm{Cov}(\mathrm{TP}, \mathrm{FN}|\bar{W}_i, Z_i) = 0$ since X_i^j of the observations are conditionally independent under Assumption 2. However, the normal density function is continuous whereas the possible values of $\bar{X}_i$ are discrete (e.g., $(0, \frac{1}{M_i}, \frac{2}{M_i}, \ldots, 1)$). Therefore, we adjust the probability density to probability mass by utilizing the half-unit continuity correction (Rumsey, 2006). Let Φ denote the cumulative distribution function of $f(\bar{X}_i|\bar{W}_i, Z_i)$:

$$\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i) = \Phi\left(\bar{X}_i + \frac{1}{M_i} \times 0.5\right) - \Phi(\bar{X}_i - \frac{1}{M_i} \times 0.5). \qquad (7)$$

Theorem 2: *Let the labeled dataset be an i.i.d random sample drawn from the population. Suppose that Assumptions 1 and 2 hold and M_i is large enough, β in Equation (4) can be approximately estimated by applying MLE to Equation (5) where $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ is estimated by Equation (7).*

The time complexity of the approximation method is much better than that of the exact method. Specifically, the time complexity of using Equation (7) to calculate the corrected probability is linear in M_i ($O(M_i)$) since we only need to calculate the normal density function once to derive $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ for each possible value of $\bar{X}_i$. However, the time complexity of using Equation (6) is quadratic in M_i ($O({M_i}^2)$) since there is an extra loop in *h*.

To apply this method in practice, we would need to decide on a threshold value of group sample size between using the solution in Theorem 1 versus Theorem 2. We borrow the rule of thumb for the normal approximation of a binomial random variable (Schader & Schmid, 1989), which states that the normal approximation is appropriate only if all the realization values within three standard deviations around $\mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)$ fall within the range of $\bar{X}_i$ (0 to 1), that is:

$$\mu_i \pm 3\sigma_i \in (0, 1).$$

Measurement Error Model 3: Approximated Solution by Law of Large Numbers

When the sample size is large enough, the mean value of any random variable may converge to its expected value. In our problem, when M_i is large enough, $\bar{X}_i$ converges to its conditional expectation $\mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)$ by the strong law of large numbers (LLN) (Feller, 2008).

$$\mathrm{P}(\bar{X}_i = \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)|\bar{W}_i, Z_i) = 1. \qquad (8)$$

In this case, we can have a simple and fast solution by rewriting Equation (5) with the following probability function:

$$\begin{aligned}\mathrm{P}(Y_i|\bar{W}_i, Z_i) &= \mathrm{P}(Y_i|\bar{X}_i = \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i), Z_i)\,\mathrm{P}(\bar{X}_i \\ &\qquad = \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)|\bar{W}_i, Z_i) \\ &= \mathrm{P}(Y_i|\bar{X}_i = \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i), Z_i). \qquad (9)\end{aligned}$$

The right side of the first equality does not have a summation and it only has the term $\bar{X}_i = \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)$, since Equation (8) implies that $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i) = 0$ for all other values of $\bar{X}_i$. The time complexity of using Equation (8) is even faster than using Equation (7) since we only need to calculate $\mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)$. However, there is no general rule in the literature for LLN approximation. To provide guidance, we follow the rule of thumb for normal approximation to derive one rule for LLN approximation. Specifically, the rule states that the law of large numbers is appropriate only if all the realization values within three standard deviations around $\mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)$ fall within the small range of $\mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i)$ (± 0.05). Empirically, researchers can adjust the small range based on their tolerance for error.

$$\mu_i \pm 3\sigma_i \in (\mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i) - 0.05, \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i) + 0.05).$$

Theorem 3. *Let the labeled dataset be the random sample i.i.d drawn from the population. Assuming that Assumptions 1 and 2 hold, β in Equation (4) can be approximately estimated by applying MLE to Equation (5), where $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ is estimated by Equation (8) when M_i is large enough.*

Full Solution of Case 1. In summary, our full solution for Case 1 involves three steps. First, we train the classifier in the predictive stage with the following criteria. The classifier produces predicted values that satisfy Assumptions 1-2. Second, for each aggregated group, we decide on one of three methods for estimation based on the rules. Finally, among the classifiers that satisfy Assumptions 1-2, we select the most precise classifier that leads to the smallest variance of the estimated regression coefficient.

Case 2: Sum of Proxy Variable as the Focal Independent Variable

This section analyzes the case where the sum of the proxy variable is utilized as the focal independent variable in the generalized linear model. Roughly speaking, the solution is conceptually the same as that in Case 1. The first two solutions are still applicable with minor adjustments. The third LLN solution is not applicable because the law of large numbers is not applicable to the sum of the proxy variable in Case 2. Due to the page limit, we omitted the technical details.

Case 3: Mean of Proxy Variable as the Dependent Variable

In this section, we analyze the case where the mean of the proxy variable is utilized as the dependent variable in the regression in the econometric stage. The dependent variable in the econometric stage is denoted by $\bar{Y}_i$, which is the mean of the true variable for the aggregated group *i*. M_i denotes the number of observations for aggregating $\bar{Y}_i$. Y_i^j denotes the true label of the observation j for aggregating $\bar{Y}_i$. For example, Deng et al. (2018) examined the impact of stock returns on microblog sentiment. They defined the stock-level "sentiment score" ($\bar{Y}_i$) by the mean of the microblog-level labels (label being "positive" or "non-positive") within a specific day *i*. M_i was the number of microblogs on that day.

True Model in the Econometric Stage

For the true regression model in Case 3, since the dependent variable is a proportion variable ranging from 0 to 1 based on discrete outcomes, the most rigorous specification should utilize the beta-binomial model to fit this distribution (Martin et al., 2020; Oberhofer & Pfaffermayr, 2014). This section also serves the purpose of illustrating the generalizability and applicability of our proposed solution on different types of probabilistic models, in addition to GLM in Cases 1 and 2.

It is intuitive to assume that the number of positive labels follows a binomial distribution:

$M_i\bar{Y}_i \sim \mathrm{Binomial}(M_i, p_i)$, and $\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_i^j$, $i = 1, \dots, N$,

where p_i is the probability of $Y_i^j = 1 =$ "positive" and can also be interpreted as the expectation of Y_i^j. p_i follows a beta distribution. Formally:

$$p_i \sim \mathrm{Beta}(\theta_i\varphi_i, (1-\theta_i)\varphi_i).$$

The beta distribution was chosen because of its wide applicability to popular special cases, such as uniform distribution. There are two parameters of a beta distribution. The first parameter θ_i represents the expectation of this distribution, $\mathrm{E}(p_i) = \theta_i$. The second parameter φ_i acts as a scaling factor that affects the variance of the distribution. A larger φ_i implies less variance.

In the literature, researchers also assume the logistic regression function between θ_i and X_i:

$$\theta_i = \mathrm{E}(p_i|X_i) = \mathrm{E}(\bar{Y}_i|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}. \tag{10}$$

Given these model setups, $\mathrm{P}(\bar{Y}_i|X_i)$ follows the probability mass function of the beta-binomial distribution:[12]

$$\mathrm{P}(\bar{Y}_i|X_i) = \binom{M_i}{M_i\bar{Y}_i}\frac{\mathrm{B}(\theta_i\varphi_i + M_i\bar{Y}_i,\ (1-\theta_i)\varphi_i + M_i - M_i\bar{Y}_i)}{\mathrm{B}(\theta_i\varphi_i, (1-\theta_i)\varphi_i)}. \tag{11}$$

[12] In Equation (11), M_i is included in X_i to simplify the notation. In hybrid studies, researchers may or may not consider M_i as a predictor for $\bar{Y}_i$ in Equation (10). Our model can be applied to both cases.

Table 3. Confusion Matrix for Dependent Variable

	$Y_i^j = 1$	$Y_i^j = 0$	Sum
$y_i^j = 1$	h (TP)	$M_i\bar{y}_i - h$ (FP)	$M_i\bar{y}_i$
$y_i^j = 0$	$M_i\bar{Y}_i - h$ (FN)	$(M_i - M_i\bar{Y}_i) - (M_i\bar{y}_i - h)$ (TN)	$M_i - M_i\bar{y}_i$
Sum	$M_i\bar{Y}_i$	$M_i - M_i\bar{Y}_i$	M_i

Measurement Error Model 1: Exact Solution by Binomial Distribution

However, in hybrid studies, Y_i^j is unobservable and the proxy variable y_i^j is observable for all records in both labeled and unlabeled sets. Similar to Cases 1 and 2, when researchers ignore the measurement error of Y_i^j, they estimate the following regression model:

$$\mathrm{P}(\bar{y}_i|X_i) = \binom{M_i}{M_i\bar{y}_i}\frac{\mathrm{B}(\theta_i\varphi_i + M_i\bar{y}_i,\ (1-\theta_i)\varphi_i + M_i - M_i\bar{y}_i)}{\mathrm{B}(\theta_i\varphi_i, (1-\theta_i)\varphi_i)}, \tag{12}$$

where $\theta_i = \mathrm{E}(\bar{y}_i|X_i) = \frac{\exp(X_i\hat{\beta})}{1+\exp(X_i\hat{\beta})}$. However, this probability function does not account for the error between $\bar{y}_i$ and $\bar{Y}_i$. The correct objective function for MLE should be re-written as,

$$\mathrm{P}(\bar{y}_i|X_i) = \sum_{\bar{Y}_i}\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i)\mathrm{P}(\bar{Y}_i|X_i). \tag{13}$$

The second term on the right side of Equation (13) can be computed by Equation (11), which is the true regression model. The first term is the measurement error model that captures the relationship between $\bar{y}_i$ and $\bar{Y}_i$. To estimate this model, we impose one assumption,

Assumption 3: *Define* $e_i^g = y_i^g - \mathrm{E}(y_i^g|X_i, Y_i^g)$. *We assume* e_i^g *is uncorrelated with* y_i^h *and* Y_i^h *for all* $h \neq g$.

The role of Assumption 3 is similar to Assumption 2. Assumption 3 implies that the residual term (e_i^g) does not correlate with the true labels and predicted labels of other observations. With Assumption 3, the results can be greatly simplified because we do not need to consider the values of other rows' variables when predicting y_i^g.

Theorem 4: *Let the labeled dataset be the i.i.d. random sample drawn from the population. Given Assumption 3,* β *in Equation (11) can be consistently estimated by applying MLE to Equation (13), where* $\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i)$ *can be decomposed as:*

$$\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i) = \sum_{h=0}^{M_i\bar{y}_i}\mathrm{B}(h; M_i\bar{Y}_i, \mathrm{Pr}_i^{\mathrm{TP}}) \times \mathrm{B}(M_i\bar{y}_i - h; M_i - M_i\bar{Y}_i, \mathrm{Pr}_i^{\mathrm{FP}}). \tag{14}$$

$\mathrm{B}(h; M_i\bar{Y}_i, \mathrm{Pr}_i^{\mathrm{TP}})$ is the binomial probability of "h TP observations out of $M_i\bar{Y}_i$ actual positive observations." $\mathrm{B}(M_i\bar{y}_i - h; M_i - M_i\bar{Y}_i, \mathrm{Pr}_i^{\mathrm{FP}})$ is the binomial probability of "$M_i\bar{y}_i - h$ FP observations out of $M_i - M_i\bar{Y}_i$ actual negative observations". $\mathrm{Pr}_i^{\mathrm{TP}}$ and $\mathrm{Pr}_i^{\mathrm{FP}}$ are the probabilities of TP and FP observations at the individual level, which are $\mathrm{P}(y_i^j = 1|Y_i^j = 1, X_i)$ and $\mathrm{P}(y_i^j = 1|Y_i^j = 0, X_i)$. The notations of the confusion matrix for the dependent variable are shown in Table 3. The proof of this theorem and approximated solutions for the large sample case are qualitatively similar to Case 1, which are shown in Appendix B. The full solution of Case 3 is similar to Case 1 and the details were omitted for brevity.

Case 4: Sum of Proxy Variable as the Dependent Variable

In this section, we analyze the case where the sum of the proxy variable is utilized as the dependent variable in the econometric model. The dependent variable in the econometric stage is denoted by S_{Y_i}, which is the sum of the true variable for the aggregated group *i*. Y_i^j denotes the true label of the observation *j* in group *i*. For example, Jabr et al. (2014) examined how the contribution of users who provided solutions online is affected by the recognition mechanism in the user support forums. The "contribution level" (S_{Y_i}) of each user is defined as the sum of post-level labels (label being "solution post" or not) and conceptually it should not be specified as a proportion. A proportion variable in this type of application captures the "quality" not the "quantity" of contribution.

True Model in the Econometric Stage

For the true model in the econometric stage, since X_i may influence S_{Y_i} through both $\mathrm{P}(M_i|X_i)$ and $\mathrm{P}(S_{Y_i}|X_i, M_i)$, we adopt the beta-binomial-Poisson model to model the relationship between X_i and S_{Y_i} (Lora & Singer, 2008; Zhu et al., 2003). We first model $\mathrm{P}(M_i|X_i)$ by Poisson model. Next, we model $\mathrm{P}(S_{Y_i}|X_i, M_i)$ by beta-binomial model. The Poisson model for the sample size M_i is given by

$$M_i \sim \mathrm{Poisson}(\lambda_i) \text{ and } \lambda_i = \mathrm{E}(M_i|X_i),$$
$$=> \mathrm{P}(M_i|X_i) = \frac{e^{-\lambda_i}{\lambda_i}^{M_i}}{M_i!}, \qquad \lambda_i = \exp(X_i\gamma). \tag{15}$$

Given M_i, we assume the sum of the labels follows a beta-binomial distribution,

$$S_{Y_i} \sim \mathrm{Binomial}(M_i, p_i), p_i \sim \mathrm{Beta}(\theta_i\varphi_i, (1-\theta_i)\varphi_i), \text{and } S_{Y_i} = \sum_{j=1}^{M_i} Y_i^j, i = 1, \ldots, N.$$
$$=> \mathrm{P}(S_{Y_i}|X_i, M_i) = \binom{M_i}{S_{Y_i}} \frac{\mathrm{B}(\theta_i\varphi_i + S_{Y_i},\ (1-\theta_i)\varphi_i + M_i - S_{Y_i})}{\mathrm{B}(\theta_i\varphi_i, (1-\theta_i)\varphi_i)}, \theta_i = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)},$$

where $\theta_i = \mathrm{E}(\bar{Y}_i|X_i)$. Then, we can derive the relationship between S_{Y_i}, θ_i, and λ_i as follows,

$$\mathrm{E}(S_{Y_i}|X_i) = \mathrm{E}\big(\mathrm{E}(S_{Y_i}|X_i, M_i)|X_i\big) = \mathrm{E}(M_i\theta_i|X_i) = \theta_i\lambda_i.$$

Therefore, to investigate the effect of X_i on S_{Y_i}, we need to estimate λ_i and θ_i, which denote the expectation of M_i and expectation of $\bar{Y}_i$ respectively.

Measurement Error Model

For $\mathrm{P}(M_i|X_i)$, researchers can directly estimate it by utilizing the Poisson model since all the variables are precisely measured. For $\mathrm{P}(S_{Y_i}|X_i, M_i)$, the true variable Y_i^j is unobservable and only y_i^j is observable for all records. If we estimate this regression model with S_{y_i} as the dependent variable, the estimated β is inconsistent.

To eliminate the inconsistency, researchers can follow the solution in Case 3, where the true model specification is the beta-binomial model. Therefore, the solution in Case 3 is part of Case 4. Finally, we can derive the marginal effect of X_{iq} by

$$\frac{\partial\mathrm{E}(S_{Y_i}|X_i)}{\partial X_{iq}} = \gamma_q \exp(X_i\gamma)\,\Lambda(X_i\beta) + \exp(X_i\gamma)\,\Lambda'(X_i\beta)\beta_q,$$

where X_{iq} denotes the q_{th} independent variable, $\Lambda()$ denotes the logistic cumulative distribution function, $\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$, and $\Lambda'()$ is the derivative of $\Lambda()$, $\Lambda()(1-\Lambda())$. Moreover, since $\mathrm{P}(M_i|X_i)$ is separately estimated, researchers can utilize other count models, such as the negative binomial model (Lora & Singer, 2011).

Simulation Results

This section reports the simulation results for assessing the effectiveness of the proposed methods. In the predictive stage, we use a real-world dataset and there is no simulated data. Simulation is only employed at the econometric stage so that we know the true values of the regression coefficients.

In the first stage, the real-world dataset is about predicting online review ratings by textual review contents for "Musical Instruments" on Amazon (He & McAuley, 2016). The original ratings for each review had five categories and we generated a binary label, "sentiment," by relabeling rating value 5 as positive and ratings 1-4 as negative. We compiled two datasets with different numbers of reviews per product. The first dataset included 87,339 reviews for 10,000 products. The number of reviews per product ranged from 1 to 1066, with a standard deviation of 31.424. Since 8467 products had few reviews (less than 10), this dataset was considered a small sample case that better fit the exact solution approach. The second dataset also contained 10,000 products and was constructed for the large sample case that better fit the two approximated solutions. In the second dataset, 6,000 products had less than 10 reviews, whereas 4,000 products had more than 100 reviews. In total, the second dataset included around 450,000 reviews. As a result, the mean or sum of the labels of those 4,000 products met the requirement of the approximated solution by the normal distribution. Moreover, the "large sample" here refers to the number of reviews per product, not the number of products. Both datasets comprised around 60% positive reviews and 40% negative reviews. The classification method was a state-of-art algorithm, XGBoost (Chen & Guestrin, 2016). We utilized the RTextTools package in R to generate the term frequency matrix of the reviews and preprocess the reviews by stemming words, converting the text to its lowercase, and removing the punctuation, stop words, and numbers. We also removed sparse terms that occurred in less than 2% of the reviews.

Regarding the evaluation procedure, we randomly sampled 17,547 rows for the labeled dataset and the remaining dataset was used for the unlabeled set, although we knew the actual labels of all rows in both datasets. Because of this property, we could evaluate the second-stage regression bias by using true labels versus predicted labels. Next, we built the XGBoost classifier on the labeled set and used that classifier to predict the labels on the unlabeled set, pretending that we did not know the labels on the unlabeled set. We also used fivefold cross-validation to compute the performance metrics of XGBoost. Predicted labels were aggregated at the product level and used in the econometric stage's regression. The true values of the aggregated variable were used in the regression as the first-best benchmarking case.

To assess the performance sensitivity of our method with respect to the classification error, we needed to create similar classifiers with different prediction performances as the candidate classifiers. Our theories suggest that the estimated coefficient is consistent for all classifiers, including classifiers with poor prediction accuracy. For this section, we tuned the "nrounds" hyper-parameter of XGBoost, which indicates the number of boosted trees. We changed "nrounds" from 4 to 200 with a step value of 4 to construct 50 classifiers with different accuracies. The prediction performances of 50 classifiers are depicted in Figure 1, where the *x*-axis is the "nrounds" divided by 4, and the *y*-axis is the performance metric value. We report the values of Accuracy, Kappa, F1, and AUC. Figure 1 suggests that the prediction performance generally becomes better as the "nrounds" value increases within 200.

In the second stage, we aggregated the proxy variable, as in the four cases described in the Theoretical Solutions section. The second-stage simulation procedure and results are discussed in the following sections. All experiments were conducted on a PC with Intel i7-6700 3.4GHz CPU and 8GB RAM.

Case 1: Mean of Proxy Variable as the Focal Independent Variable

The true econometric model is specified as the following logistic regression model with the coefficients being 1 for both right-side variables,

$$\mathrm{P}(Y_i = 1|\bar{X}_i, Z_i) = \frac{\exp(0.2 + \bar{X}_i + Z_i)}{1 + \exp(0.2 + \bar{X}_i + Z_i)}, \qquad (16)$$

where $\bar{X}_i$ is the mean value of review sentiment computed from the first stage. M_i is the number of reviews. Z_i is a simulated control variable. Among these three variables, only M_i is from the actual dataset and the other two were simulated. The simulation procedure involved several steps. We started by simulating Z_i by a normal distribution with a mean of 0 and a variance of 0.25. Given the value of Z_i, we specified $\mathrm{P}(X_i^j = 1|Z_i)$ as follows:

$$\mathrm{P}\left(X_i^j = 1|Z_i\right) = \frac{\exp(0.2 + Z_i)}{1 + \exp(0.2 + Z_i)}.$$

Third, given the conditional probability function of X_i^j, we simulated the realized value of each X_i^j by Bernoulli distribution. The sum of X_i^j followed a binomial distribution with parameters (M_i , $\mathrm{P}\left(X_i^j = 1|Z_i\right)$) . Fourth, given the simulated value of each X_i^j, if the value was 1, we randomly matched that row with a positive review record in the actual dataset. This allowed us to compute $\bar{X}_i$ easily and derive $\mathrm{P}(Y_i = 1|\bar{X}_i, Z_i)$ using Equation (16). Finally, the Bernoulli distribution was utilized to simulate Y_i using Equation (16).

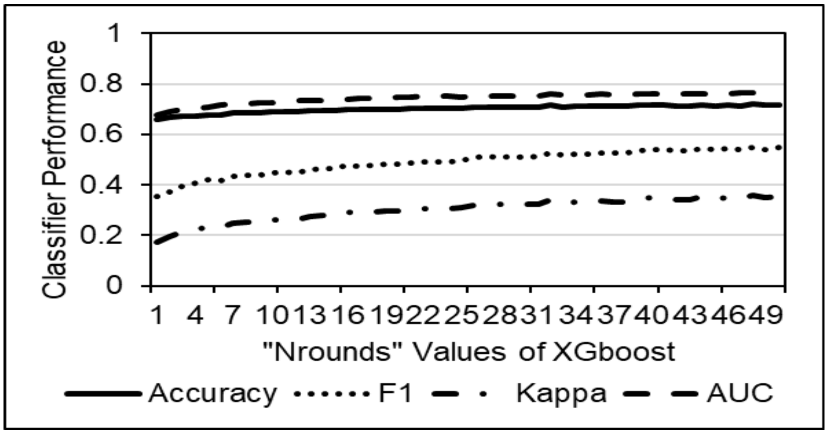
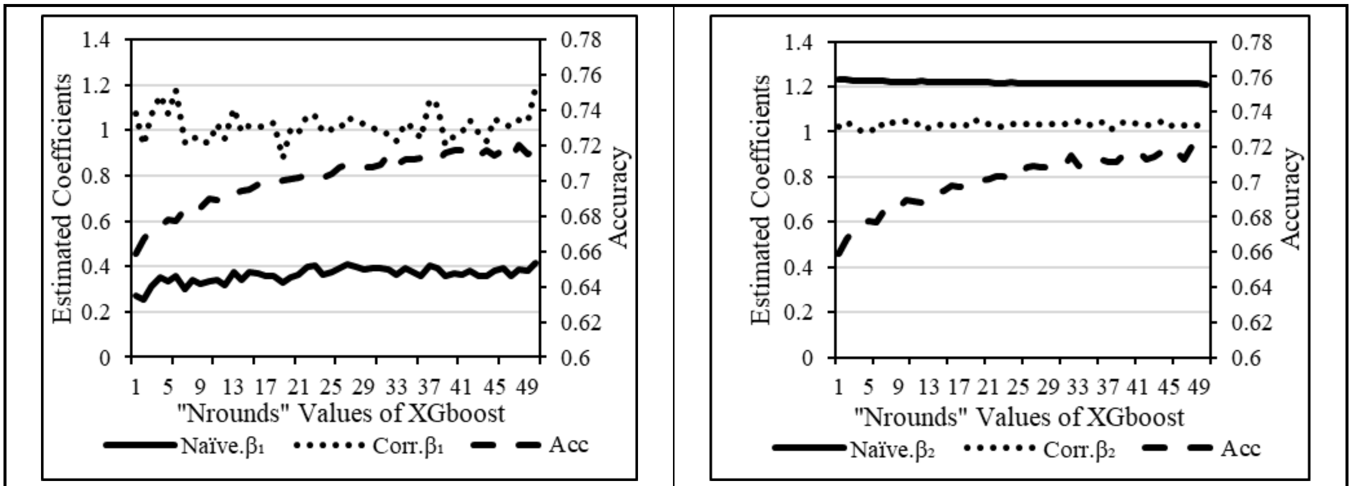


Figure 1. Classification Performance



(a) Results for focal variable (ground truth of $\boldsymbol{\beta_1}$ is 1)

(b)Results for control variable (ground truth of $\boldsymbol{\beta_2}$ is 1)

Note: $\mathrm{Naïve}.\beta_1$ refers to the coefficient of $\bar{X}_i$ without correction; $\mathrm{Corr}.\beta_1$ refers to the coefficient of $\bar{X}_i$ after using the exact solution. β_2 refers to the coefficient of Z_i. Acc refers to the classifier accuracy.

Figure 2. Simulation Results for Case 1 (Small Sample Case)

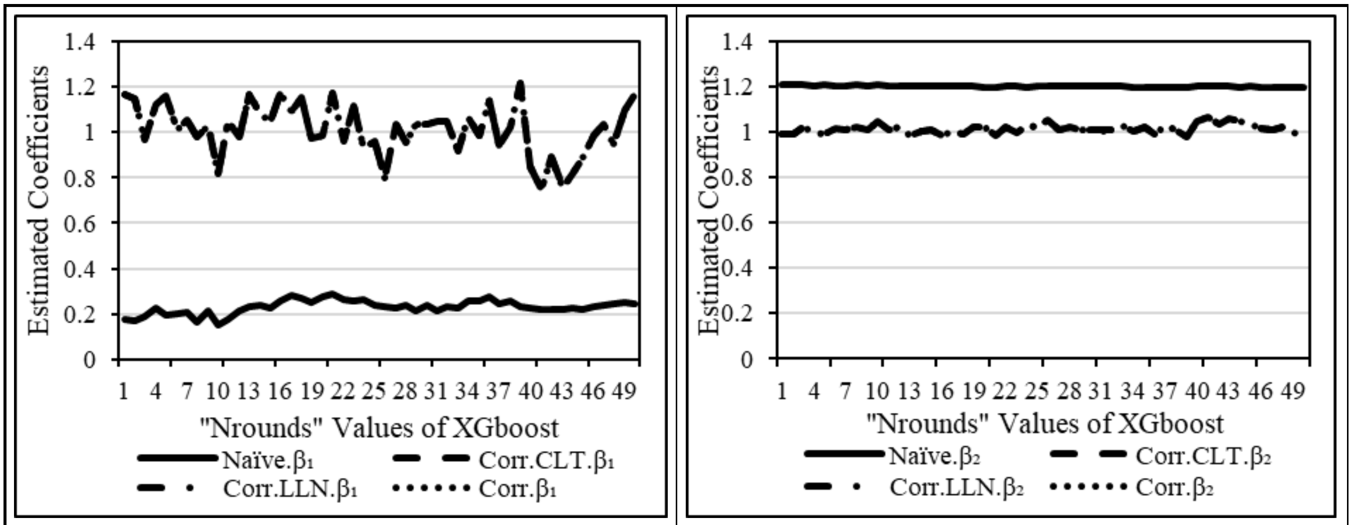
Evaluation of the Exact Solution for Small Sample Case

We first evaluated the exact solution for the small sample case. When using the true mean of review sentiment ($\bar{X}_i$) to derive the empirical regression coefficients, the estimated coefficients for $\bar{X}_i$ and Z_i were 1.094 and 1.006, respectively. This is the ideal solution and is not feasible in practice unless the text classification is perfect.

Next, we derived the regression results from classifiers with errors. For each “nrounds” value, we trained and built a classifier and derived the proxy variable W for X . The baseline case was regression without any correction: regression by using the mean of the proxy variable ($\bar{W}_i$) as the independent variable. This is also the method widely used in existing literature. Results are indicated in Figures 2a and 2b by the solid lines after we repeated the regression by using 50 different proxy variables based on different “nrounds” values. In the figures, the *x*-axis is the “nrounds” value divided by 4, and the *y*-axis on the left side is the coefficient value. The *y*-axis on the right side shows the accuracy value. The accuracy results are reported by dashed lines. We observe that the focal regression coefficient without any correction is much smaller than the true value 1. The results using Equation (5) and Equation (6) to derive the corrected coefficients for $\bar{X}_i$ and Z_i are depicted as the dotted lines in Figures 2a and 2b. These results suggest that our method corrected the estimation error and is not sensitive to the classifier performance.

Evaluation of the Approximated Solutions for Large Sample Case

We then evaluated our methods for the large sample case. We first conducted logit regression using the true mean of review sentiment ($\bar{X}_i$). The estimated coefficients for $\bar{X}_i$ and Z_i were 0.937 and 1.020. Next, we applied three correction methods to this dataset. In the first method, we applied Equation (5) and Equation (6) (exact solution) to the observations aggregated by less than 10 reviews and applied Equation (7) (CLT approximation) to the observations aggregated by over 100 reviews. However, in the second method, we applied Equation (8) (LLN approximation) to the observations aggregated by over 100 reviews. In the third method, we applied the exact solution to the whole dataset to derive the corrected coefficients. The results for the three methods are shown as the dashed, dash-dotted, and dotted lines in Figure 3. We observe that these three lines are indistinguishable. The overlap of the corrected coefficients from the three methods indicates the effectiveness of our approximated solutions when the number of reviews per product is large enough. The results also confirm that our exact and approximated solutions can indeed correct the estimation inconsistency. Moreover, the relative running time of LLN approximation, CLT approximation, and the exact solution was around 1:1.4:6.5. Specifically, the LLN approximation was the fastest. We summarize the time analysis of the first three sections in Table 4.



(a) Results for focal variable (ground truth of $\boldsymbol{\beta_1}$ is 1)

(b) Results for control variable (ground truth of $\boldsymbol{\beta_2}$ is 1)

Note: $\text{Naïve}.\beta_1$ refers to the coefficient of $\bar{X}_i$ without correction; $\text{Corr.CLT}.\beta_1$ refers to the coefficient of $\bar{X}_i$ after applying the CLT approximation solution. $\text{Corr.LLN}.\beta_1$ refers to the coefficient of $\bar{X}_i$ after applying the LLN approximation solution. $\text{Corr}.\beta_1$ refers to the coefficient of $\bar{X}_i$ after using our exact solution. β_2 refers to the coefficient of Z_i.

Figure 3. Simulation Results for Case 1 (Large Sample Case)

Table 4. CPU Run Time Analysis (minutes)

	Exact	**CLT**	**LLN**
Complexity	$\boldsymbol{O(M_i^2)}$	$\boldsymbol{O(M_i)}$	$\boldsymbol{O(1)}$
CPU run time for Case 1	32.363	7.085	5.107
CPU run time for Case 2	48.989	14.830	None
CPU run time for Case 3	54.801	24.834	15.509

Case 2: Sum of Proxy Variable as the Focal Independent Variable

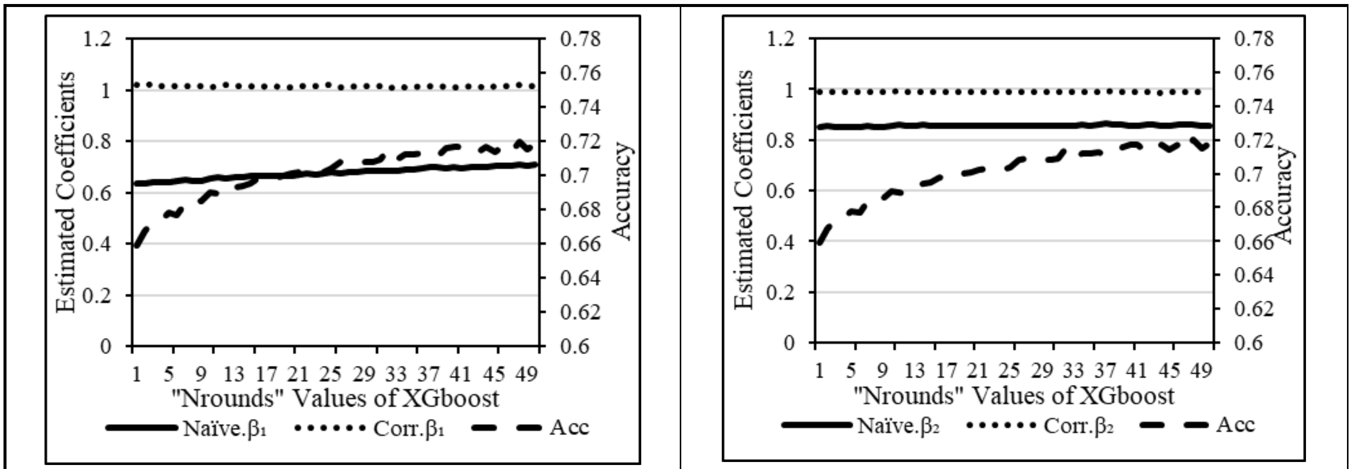
We simulated data using the following logistic regression model with coefficients of 1 for both right-side variables in the econometric stage:

$$\mathrm{P}\left(Y_i = 1|S_{X_i}, Z_i\right) = \frac{\exp\left(-2 + S_{X_i} + Z_i\right)}{1 + \exp\left(-2 + S_{X_i} + Z_i\right)}, \tag{17}$$

where $S_{X_i} = \sum_{j=1}^{M_i} X_i^j$. M_i is the total number of reviews. M_i is from the actual dataset and S_{X_i} was simulated by randomly sampling M_i reviews from the actual dataset. S_{X_i} is the sum of review sentiment. Z_i is a control variable and simulated by $Z_i = -0.5 \times S_{X_i} + \varepsilon_i$. ε_i follows the normal distribution with a mean of 0 and a standard error of 5. Given the values of S_{X_i} and Z_i, the values of $\mathrm{P}(Y_i = 1|S_{X_i}, Z_i)$ were computed by Equation (17) accordingly. Finally, the Bernoulli distribution was utilized to generate Y_i.

Evaluation of the Exact Solution for Small Sample Case

In the first-best regression analysis, the empirical coefficients for S_{X_i} and Z_i were 0.992 and 0.983, respectively. Similar to Case 1, we also constructed 50 classifiers. For each classifier, we ran the naïve regression by using the sum of the proxy variable (S_{W_i}) as the focal independent variable, and the estimated coefficient was much smaller than 1. Next, we derived the corrected coefficients for S_{X_i} and Z_i. Figure 4a and Figure 4b show that our method can produce consistent estimations for almost all classifiers.



(a) Results for Focal Variable (Ground Truth of β_1 is 1)

(b) Results for Control Variable (Ground Truth of β_2 is 1)

Note: The legend entries are the same as Figure 2.

Figure 4. Simulation Results for Case 2 (Small Sample Case)

Evaluation of the Approximated Solutions for Large Sample Case

In the first-best regression analysis, the empirical coefficients for S_{X_i} and Z_i were 0.938 and 0.954. Naïve regression again produced biased results. In contrast to the Case 1 Results section, we only had two but not three solutions because LLN was not applicable to the sum of sentiment scores. First, we applied the exact solution to the observations aggregated by less than 10 reviews and applied CLT approximation to the observations aggregated by more than 100 reviews. In the second method, we applied the exact solution to the whole dataset to derive the corrected coefficients. The results are shown as dashed lines and dotted lines in Figure 5, respectively. The figures show that the dashed lines and dotted lines are almost identical, which implies the correction effectiveness of the CLT approximation solution. Moreover, the relative running time of CLT approximation and the exact solution was around 1:3.3, indicating that utilizing CLT approximation can achieve both speed and estimation accuracy.

Case 3: Mean of Proxy Variable as the Dependent Variable

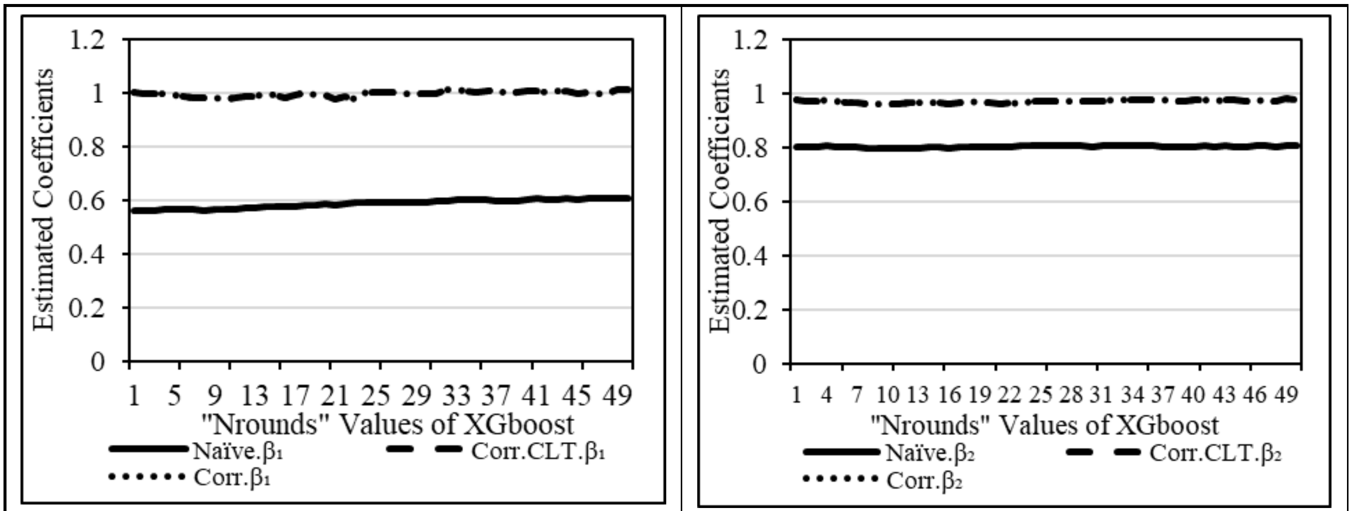
Because $\bar{Y}_i$ is a ratio between 0 and 1, we simulated the regression model using a beta-binomial regression model with coefficients of 1 for both right-side variables:

$$\mathrm{P}(\bar{Y}_i|X_i, Z_i) = \binom{M_i}{M_i\bar{Y}_i}\frac{\mathrm{B}(\theta_i\varphi_i + M_i\bar{Y}_i,\ (1-\theta_i)\varphi_i + M_i - M_i\bar{Y}_i)}{\mathrm{B}(\theta_i\varphi_i, (1-\theta_i)\varphi_i)}, \tag{18}$$

where $\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_i^j$ and $\theta_i = \mathrm{E}(\bar{Y}_i|X_i, Z_i) = \frac{\exp(0.5+X_i+Z_i)}{1+\exp(0.5+X_i+Z_i)}$. $\bar{Y}_i$ is the mean of review sentiment. M_i is the number of reviews from the actual dataset. For the simulation process, we started by simulating X_i using a standard normal distribution. Second, we simulated Z_i by $Z_i = -0.5 \times X_i + \varepsilon_i$, where ε_i followed a standard normal distribution. Third, given the values of X_i and Z_i, we derived $\mathrm{E}(\bar{Y}_i|X_i, Z_i)$ using $\frac{\exp(0.5+X_i+Z_i)}{1+\exp(0.5+X_i+Z_i)}$. Fourth, we set φ_i as 4, which implied that the correlation between binary observations within the aggregated group was 0.2. Given the simulated θ_i and the value of φ_i, $\mathrm{P}(\bar{Y}_i|X_i, Z_i)$ was computed accordingly by Equation (18) and $M_i\bar{Y}_i$ was generated by beta-binomial distribution. Finally, we randomly matched records in actual review data with the simulated X_i and Z_i, where the number of reviews and the sum of review sentiment equalled M_i and $M_i\bar{Y}_i$.

Evaluation of the Exact Solution for Small Sample Case

In the first-best scenario, the empirical coefficients for X_i and Z_i were 1.037 and 1.004. We ran the naïve regression by using the mean of the proxy variable ($\bar{y}_i$) as the dependent variable, following Equation (12). Next, we used Equation (13) and Equation (14) to derive the corrected coefficients for X_i and Z_i. Figures 6a and 6b show that the exact solution corrected the coefficient inconsistency and was not very sensitive to the classifier performance, whereas the naïve solution's performance was poor across all classifiers.

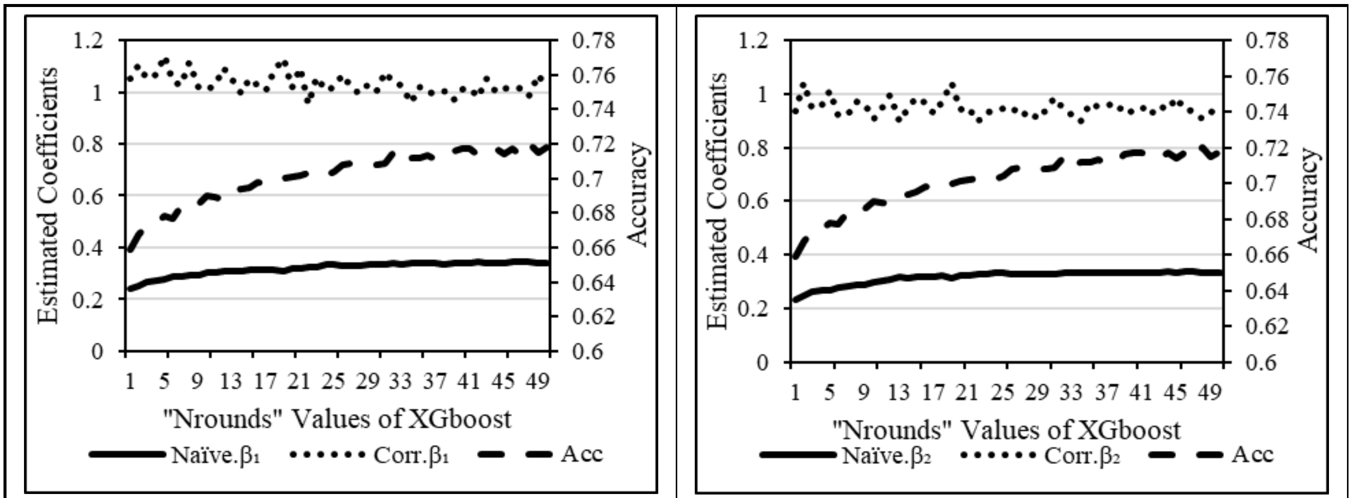


(a) Results for Focal Variable (Ground Truth of β_1 is 1)

(b) Results for Control Variable (Ground Truth of β_2 is 1)

Note: The legend entries are the same as Figure 3.

Figure 5. Simulation Results for Case 2 (Large Sample Case)



(a) Results for Focal Variable (Ground Truth of β_1 is 1)

(b) Results for Control Variable (Ground Truth of β_2 is 1)

Note: The legend entries are the same as Figure 2.

Figure 6. Simulation Results for Case 3 (Small Sample Case)

Evaluation of the Approximated Solutions for Large Sample Case

In the first-best scenario, the empirical coefficients for X_i and Z_i were 1.010 and 0.997, respectively. Next, we applied three correction methods to the dataset. In the first correction method, we applied Equation (13) and Equation (14) (exact solution) to the observations aggregated by less than 10 reviews and applied the CLT approximation solution to the observations aggregated by more than 100 reviews. The results are depicted as dashed lines in Figures 7a and 7b. In the second correction method, we applied the LLN approximation solution to the observations aggregated by more than 100 reviews. The results are shown as the dotted lines in Figures 7a and 7b. In the third correction method, we applied Equation (13) and Equation (14) (exact solution) to the whole dataset to derive the corrected coefficients. The results are shown as dash-dotted lines in Figures 7a and 7b. The results show that the CLT approximation and the exact solution obtained almost the same coefficients, which indicates the effectiveness of the CLT approximation solution. However, the LLN approximation solution only partially corrected the estimation inconsistency, possibly because the sample size for aggregation was not large enough. The relative running time of LLN approximation, CLT approximation, and the exact solution was 1:1.6:3.7, which shows the superior speed of the approximated solutions.

Case 4: Sum of Proxy Variable as the Dependent Variable

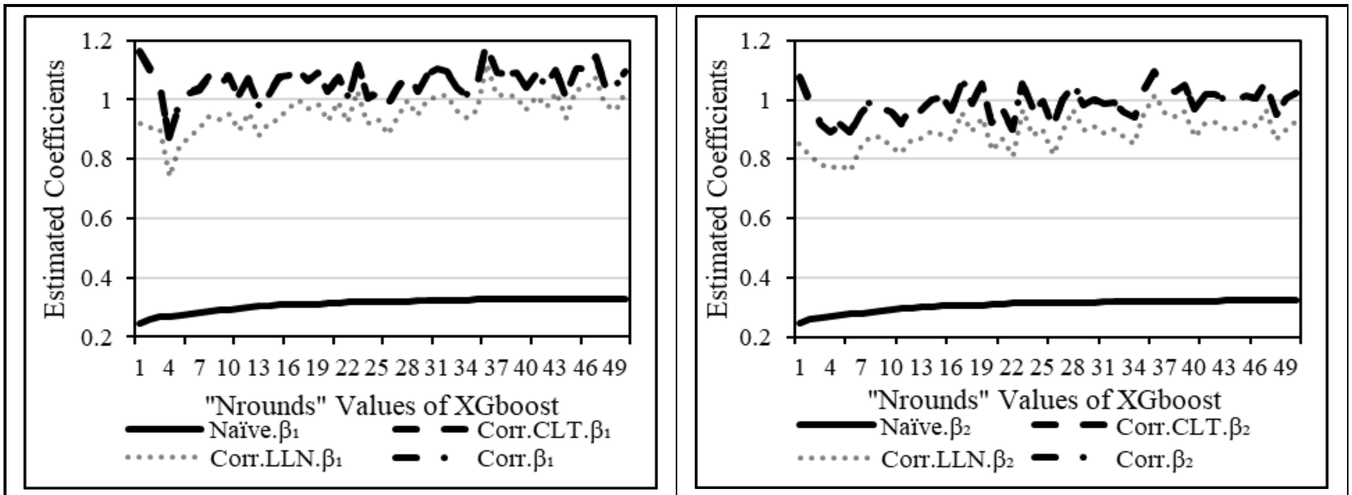
We simulated data using the following model with two right-side variables,

$$\mathrm{P}(M_i|X_i, Z_i) = \frac{{\lambda_i}^{M_i} e^{-\lambda_i}}{M_i!},\ \lambda_i = \exp(1.5 + X_i + Z_i), \tag{19}$$

$$\mathrm{P}(S_{Y_i}|M_i, X_i, Z_i) = \binom{M_i}{S_{Y_i}} \frac{\mathrm{B}\left(\theta_i\varphi_i + S_{Y_i},\ (1-\theta_i)\varphi_i + M_i - S_{Y_i}\right)}{\mathrm{B}\left(\theta_i\varphi_i, (1-\theta_i)\varphi_i\right)}, \tag{20}$$

where $\theta_i = \frac{\exp(0.5+X_i+Z_i)}{1+\exp(0.5+X_i+Z_i)}, \varphi_i = 4, Z_i = -0.5 \times X_i + \varepsilon_i$, and $S_{Y_i} = \sum_{j=1}^{M_i} Y_i^j$. S_{Y_i} is the sum of review sentiment. M_i is the number of reviews, which was simulated based on Equation (19). X_i and Z_i were simulated following the same procedure in Case 3. The simulation process for Case 4 had two stages. The first stage was used to simulate M_i using the Poisson model. The second stage was used to simulate S_{Y_i} using beta-binomial model. The simulation process of the second stage was the same as Case 3. For the first stage, given simulated values of X_i and Z_i, we derived $\mathrm{P}(M_i|X_i, Z_i)$ using Equation (19) and generated M_i using Poisson distribution. The equations also imply that the theoretical beta coefficients of X_i and Z_i were 1 in both models.

In this case, we needed to estimate both $\mathrm{P}(M_i|X_i, Z_i)$ and $\mathrm{P}(S_{Y_i}|M_i, X_i, Z_i)$. We utilized the dataset for the small sample case to evaluate our method. We first ran the Poisson model to estimate $\mathrm{P}(M_i|X_i, Z_i)$. The empirical coefficients for X_i and Z_i were 0.985 and 0.987, respectively. Next, we ran the beta-binomial model using the sum of true review sentiment S_{Y_i} to estimate $\mathrm{P}(S_{Y_i}|M_i, X_i, Z_i)$. The empirical coefficients for X_i and Z_i were 1.027 and 1.022, respectively. For each classifier, we ran the naïve regression by using the sum of the proxy variable (S_{y_i}) as the dependent variable. Next, we used Equation (13) and Equation (14) (exact solution) to derive the corrected coefficients for X_i and Z_i. Figures 8a and 8b show the coefficient results. All results are qualitatively the same as those in the previous three cases.

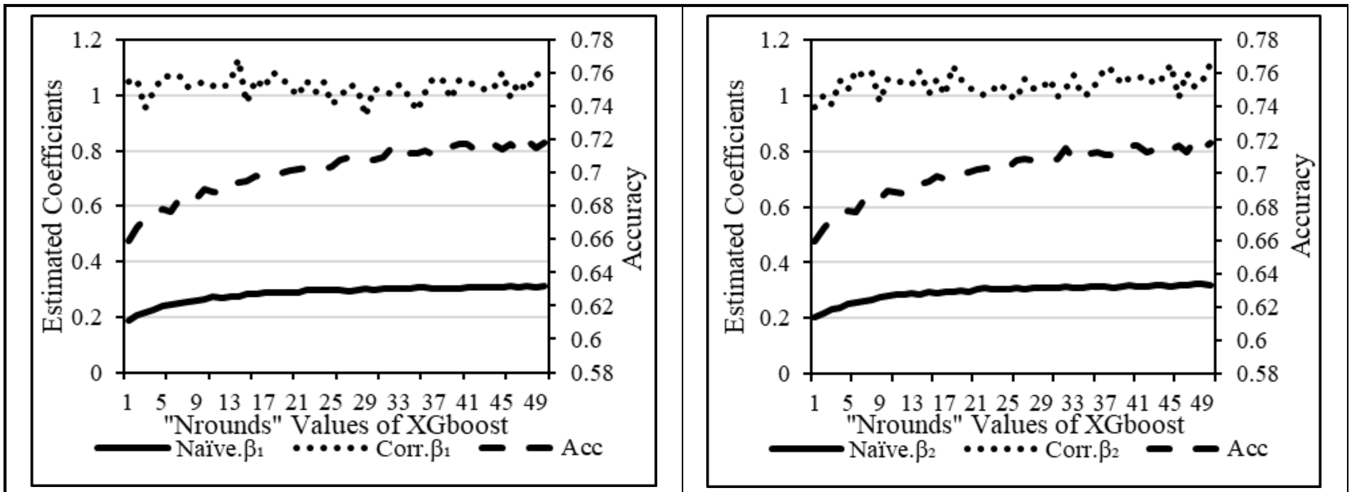


(a) Results for Focal Variable (Ground Truth of β_1 is 1)

(b) Results for Control Variable (Ground Truth of β_2 is 1)

Note: The legend entries are the same as Figure 3.

Figure 7. Simulation Results for Case 3 (Large Sample Case)



(a) Results for Focal (Ground Truth of β_1 is 1)

(a) Results for Control Variable (Ground Truth of β_2 is 1)

Note: The legend entries are the same as Figure 2.

Figure 8. Simulation Results for Case 4 (Small Sample Case)

Table 5. Robustness Results for Case 1 and Case 2

	Mean				**Sum**			
	$\mathrm{Pr}^{\mathrm{TP}}$	$\mathrm{Pr}^{\mathrm{FP}}$	**Naïve**	**Our method**	$\mathrm{Pr}^{\mathrm{TP}}$	$\mathrm{Pr}^{\mathrm{FP}}$	**Naïve**	**Our method**
Setting 1	0.600	0.400	0.336	0.854	0.600	0.400	0.694	1.148
	0.700	0.300	0.583	0.893	0.700	0.300	0.842	1.105
	0.800	0.200	0.734	0.866	0.800	0.200	0.951	1.056
	0.900	0.100	0.938	0.979	0.900	0.100	0.995	1.015
Setting 2	0.600	0.400	1.540	1.020	0.600	0.400	1.432	1.028
	0.700	0.300	1.685	1.029	0.700	0.300	1.569	1.019
	0.800	0.200	1.469	1.041	0.800	0.200	1.384	1.002
	0.900	0.100	1.198	1.010	0.900	0.100	1.176	1.008

Robustness Analysis

The validity of our correction method rests on the assumptions in the Theoretical Solutions section. For the robustness tests, we conducted experiments involving the violation of the assumptions.

Robustness Evaluation of Assumption 2 for Case 1 and Case 2

First, we simulated the sample size of the aggregated group following uniform distribution, $M_i \sim U(1,10)$ and $M_i \sim U(20,30)$ for two settings ($i = 1, \ldots, 5000$). Second, we simulated X_i^j using:

Setting 1: $\mathrm{P}\left(X_i^j = 1 \middle| X_i^{j-1}\right) = \frac{1}{1+\exp\left(-\alpha_i + X_i^{j-1}\right)}$;

Setting 2: $\mathrm{P}(X_i^j = 1 | X_i^{j-1}) = \frac{1}{1+\exp\left(-\alpha_i - X_i^{j-1}\right)}$,

where X_i^1 was simulated by Bernoulli distribution with event probability $p_i \sim U(0,1)$ and $\alpha_i = \log\left(\frac{p_i}{1-p_i}\right)$. In this case, X_i^j and X_i^{j-1} were correlated, which violates Assumption 2 that two variables are conditionally independent. Third, we modified X_i^j by setting the true positive rate ($\mathrm{Pr}^{\mathrm{TP}}$) and false positive rate ($\mathrm{Pr}^{\mathrm{FP}}$) to obtain W_i^j. Finally, we simulated the dependent variable using $Y_i = 1 + \bar{X}_i + \varepsilon_i$ and $Y_i = 1 + S_{X_i} + \varepsilon_i$. ε_i followed the standard normal distribution. In Table 5, the results show that when Assumption 2 is violated, our method can still correct the bias (closer to the correct value 1 than the naïve method). Moreover, when the classifier's accuracy becomes better, our method's performance becomes better.

Robustness Evaluation of Assumption 3 for Case 3 and Case 4

First, we simulated the sample size of the aggregated group following uniform distribution for Case 3, $M_i \sim U(1,10)$ $(i = 1, \ldots, 5000)$. For Case 4, we simulated the sample size of the group following zero-truncated Poisson distribution with the mean as $\exp(X_i + 2)$ and X_i following a standard normal distribution. Second, we computed the expectation of the mean of the labels by $\mathrm{E}(\bar{Y}_i) = \frac{1}{1+\exp\,(-X_i+0.2)}$. Third, we generated the sum of labels by beta-binomial distribution with the expectation of event probability as $\mathrm{E}(\bar{Y}_i)$ and φ_i as 1. Fourth, we modified Y_i^j by setting $\mathrm{Pr}^{\mathrm{TP}}$ and $\mathrm{Pr}^{\mathrm{FP}}$ to introduce misclassification and obtained y_i^j. Specifically, we simulate y_i^j using:

Setting 1: $\mathrm{P}(y_i^j = 1 \mid Y_i^j,\, y_i^{j-1}) = \frac{1}{1+\exp\,(-\beta \times Y_i^j - y_i^{j-1})}$,

Setting 2: $\mathrm{P}(y_i^j = 1 \mid Y_i^j,\, y_i^{j-1}, Y_i^{j-1}) = \frac{1}{1+\exp\left(-\beta \times Y_i^j - 0.5 \times y_i^{j-1} - 0.5 \times Y_i^{j-1}\right)}$.

By these two equations, the assumption that y_i^j and y_i^{j-1} (Y_i^{j-1}) are conditionally independent is violated. In the experiments, we changed the values of β to obtain different values of $\mathrm{Pr}^{\mathrm{TP}}$and $\mathrm{Pr}^{\mathrm{FP}}$. The results are reported in Table 6. The results show that when Assumption 3 is violated, our method can partially but not fully correct the bias. However, our method still outperforms the naïve method.

Applications to Real-World Data Set

In this section, we discuss the experiments conducted with a realistic second-stage regression model similar to the models published in IS literature rather than a simulated regression model as in the Simulation section. The Amazon review data described previously were utilized to conduct experiments of evaluating the solutions for Case 1 and Case 2. We only evaluated exact solutions since most products had a small number of reviews in this dataset. For the second-stage regression, we examined the impact of product review sentiment score on product sales rank. Amazon does not publicly display the total number of units sold for a product. Therefore, we used the sales rank of each product as a proxy variable for sales, following the IS literature. In Case 1, product review sentiment was operationalized as the mean of sentiment labels of all the product reviews before the sales rank was observed. We included one control variable, product price. In Case 2, product review sentiment was operationalized as the sum of sentiment labels of the product. We included two control variables, product price, and the total number of reviews. The model specifications are given below:

Case 1: $\mathrm{Log}(Rank) = \beta_1 + \beta_2\,\mathrm{Log}(Avg_{\mathrm{Sent}} + 1) + \beta_3 \mathrm{Log}(Price) + \varepsilon,$

Case 2: $\mathrm{Log}(Rank) = \beta_1 + \beta_2\,\mathrm{Log}(Sum_{\mathrm{Sent}} + 1) + \beta_3 \mathrm{Log}(Price) +$
$\beta_4 \mathrm{Log}(Num + 1) + \varepsilon.$

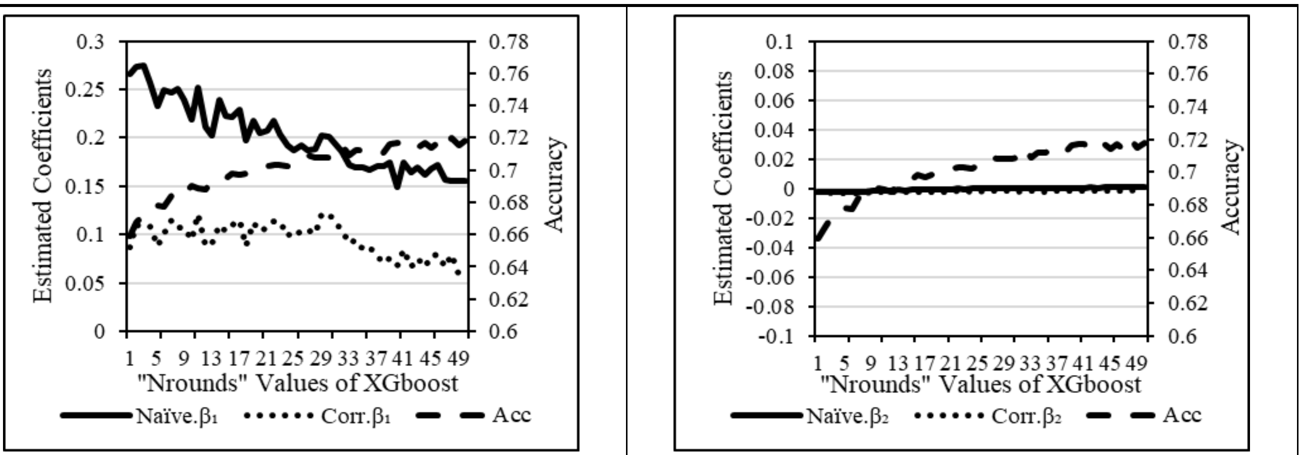
This study first conducted two regression models using true review sentiment to derive the true regression coefficients, which are reported in Table 7. The results show that the sentiment was negatively associated with the product sales rank (Lower rank means more sales). Products with more positive reviews were more likely to achieve higher sales. Furthermore, products with a higher price were more likely to obtain fewer sales. Next, similar to the Simulation section, we constructed 50 classifiers. For each classifier, we ran the naive regression by using the mean and sum of predicted sentiment as the focal independent variables (the worst case without any correction). In addition, we used our solutions to derive the corrected coefficients. Figures 9 and 10 report the results for Case 1 and Case 2, respectively. In these figures, the y-axis shows the difference between the estimated coefficients using two different methods and the true coefficient. β_1, β_2, and β_3 refer to the coefficient difference of the review sentiment, product price, and the number of reviews. The results of β_2 are almost the same in both cases. Therefore, we omitted the results of β_2 for Case 2.

Table 6. Robustness Results for Case 3 and Case 4

	Mean				Sum			
	$\mathrm{Pr}^{\mathrm{TP}}$	$\mathrm{Pr}^{\mathrm{FP}}$	Naïve	Our method	$\mathrm{Pr}^{\mathrm{TP}}$	$\mathrm{Pr}^{\mathrm{FP}}$	Naïve	Our method
Setting 1	0.708	0.507	0.183	0.798	0.735	0.521	0.288	1.001
	0.796	0.388	0.362	0.874	0.820	0.401	0.478	0.992
	0.869	0.259	0.565	0.965	0.888	0.277	0.670	1.023
	0.937	0.129	0.772	0.949	0.949	0.141	0.844	1.015
Setting 2	0.725	0.471	0.260	0.883	0.756	0.491	0.338	0.969
	0.803	0.360	0.433	0.963	0.828	0.381	0.504	0.949
	0.873	0.245	0.607	0.983	0.893	0.261	0.674	1.003
	0.939	0.121	0.775	0.963	0.949	0.136	0.833	0.981

Table 7. Estimated Regression Coefficients by Using True Review Sentiments

	Sentiment	*SD*	Price	*SD*	No. of reviews	*SD*
Case 1	-0.681	(0.047)	0.187	(0.008)		
Case 2	-0.381	(0.022)	0.174	(0.006)	-0.542	(0.022)

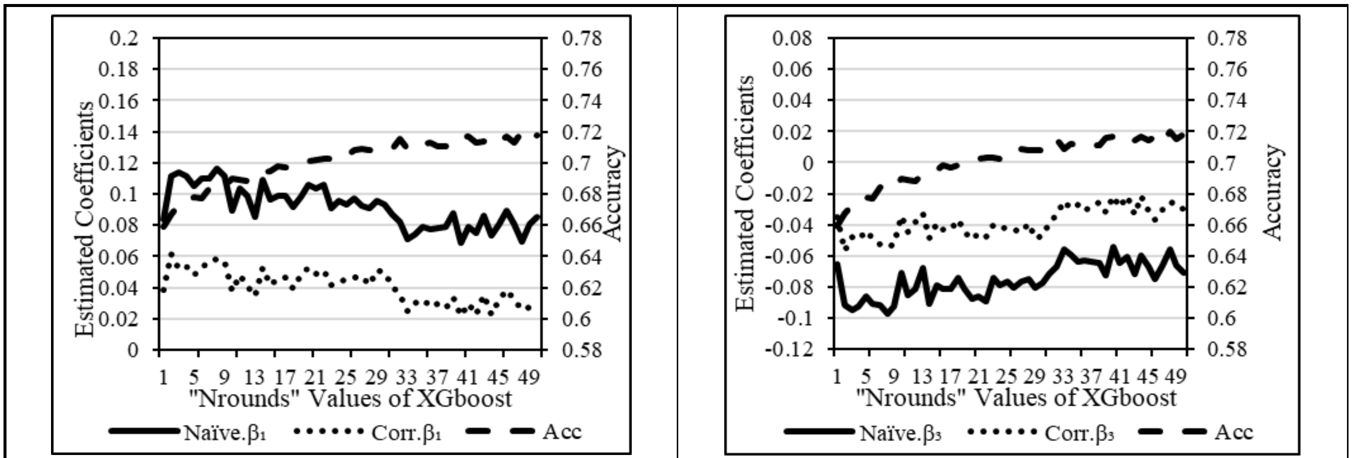


(a) Results for Review Sentiment

(b) Results for Product Price

Note: The legend entries are the same as Figure 2.

Figure 9. Coefficient Results for Case 1



(a) Results for Review Sentiment

(b) Results for No. of Reviews

Note: The legend entries are the same as Figure 2.

Figure 10. Coefficient Results for Case 2

Comparing Naïve.β_1 with 0 indicates a substantial bias in the estimator for both scenarios in the absence of any corrective method. In contrast, an analysis of Naïve.β_2 versus 0 reveals that misclassification in the sentiment variable did not bias the coefficients of the product price in either case. The reason may be that the correlation coefficients between product price and both the mean and sum of review sentiment were very small. The coefficients were only 0.006 and -0.021 with *p*-values of 0.542 and 0.040. Comparing Naïve.β_3 with 0 shows that the misclassification in the sentiment variable biased the coefficients of the total number of reviews. Comparing Corr.β_1 and Corr.β_3 with 0 underscores the efficacy of our methodology in rectifying the estimation bias.

Discussion and Conclusion

Inspired by the pioneering work of Yang et al. (2018) and the solution proposed by Qiao and Huang (2021) for individual hybrid studies, our paper proposes a new solution for aggregate-level hybrid studies, which has not been analyzed in the literature. The contribution of our study is twofold. First, we analyzed how the regression estimation bias may vary with classification error in the prediction stage and the sample size of the aggregated group. Second, our proposed solution can improve the estimation accuracy in aggregate-level hybrid studies, which seem to be more prevalent than individual-level hybrid studies. In this study, we derived theoretical formulas of consistent estimators of regression coefficients for four common types of aggregate-level hybrid studies. We evaluated our solutions using both simulation and real-world data analysis and our experimentation shows that our method can produce consistent regression estimators and that the estimation results are not very sensitive to classification error. The simulation and real-world application both show that our method can indeed correct the inconsistency of estimated coefficients by the naïve regression approach.

However, our proposed method may not achieve perfect correction performance if the three assumptions of our method are not satisfied. In particular, Assumptions 2 and 3 are more likely to be violated in practice. It is strongly advised that researchers endeavor to train classifiers that adhere closely to Assumptions 1-3 to ensure that the regression estimators are consistent. Given that the proposed method is more complicated than the naïve method, we recommend that researchers use the naïve approach in the following three scenarios. First, in cases where classification performance is exceedingly high (e.g., 95% or greater), the measurement error after aggregation is likely to be minimal, thereby diminishing the utility of our correction technique. Second, if empirical data indicate a breach of Assumptions 1-3, our method cannot guarantee the production of a consistent estimator. Under such circumstances, our method should be regarded as a supplementary test for the specific hybrid study. Lastly, when the mean of the proxy variable is used in regression and the sample size of aggregation is large, because of the law of large numbers, naïve methods (without any correction) with an unbiased classifier can produce consistent estimates. The importance of an unbiased classifier in the predictive stage deserves more attention from researchers engaged in hybrid studies.

The current study's theoretical contributions are confined to certain econometric models that can be estimated by maximum likelihood estimation (MLE), such as generalized linear models, survival models, and beta-binomial models. These findings do not extend to more complex regressions, such as panel, time series, and vector autoregression models. Theoretical results of other models could be derived using a similar probabilistic approach but it is not possible to cover all cases in one paper. Second, while our approach can accommodate multi-class outputs at the predictive stage, such an extension would require further exploration of central limit theorem-based approximations due to the computational intensity of the exact solution. Third, our results may not be applicable to other aggregation functions, such as the standard deviation of the proxy variable. Fourth, when the prediction performance is poor, the variance of regression coefficient estimates may be significantly inflated despite the estimator consistency. Therefore, it is necessary to investigate the precise interplay between coefficient variance and classification accuracy. Lastly, determining the most effective data annotation strategy for enhanced regression estimation in hybrid studies remains a pertinent topic for future investigation.

Acknowledgments

The authors thank the senior editor, the associate editor, and the anonymous reviewers for their constructive comments and valuable suggestions. This research/project was supported by the National Research Foundation, Singapore under its Industry Alignment Fund—Pre-Positioning (IAF-PP) Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Aggarwal, R., & Singh, H. (2013). Differential influence of blogs across different stages of decision making: The case of venture capitalists. *MIS Quarterly*, *37*(4), 1093-1112. https://doi.org/10.25300/MISQ/2013/37.4.05

Bentkus, V. (2005). A Lyapunov-type bound in Rd. *Theory of Probability & Its Applications*, *49*(2), 311-323. https://doi.org/10.1137/S0040585X97981123

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420066586

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., & Stefanski, L. A. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman & Hall/CRC. https://doi.org/10.1201/9781420010138

Chan, J., & Wang, J. (2018). Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science*, *64*(7), 2973-2994. https://doi.org/10.1287/mnsc.2017.2756

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, *36*(4), 1165-1188. https://doi.org/10.2307/41703503

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.293978

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, *89*(428), 1314-1328. https://doi.org/10.1080/01621459.1994.10476871

Deng, S., Huang, Z. J., Sinha, A. P., & Zhao, H. (2018). The interaction between microblog sentiment and stock return: An empirical examination. *MIS Quarterly*, *42*(3), 895-918. https://doi.org/10.25300/MISQ/2018/1426

Feller, W. (2008). *An introduction to probability theory and its applications* (Vol. 2). John Wiley & Sons. https://doi.org/10.1080/00224065.1970.11980411

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons. https://doi.org/10.1002/9780470316665

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498-1512. https://doi.org/10.1109/TKDE.2010.188

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, *31*(3), 493-520. https://doi.org/10.1287/mksc.1110.0700

Greene, W. H. (2012). *Econometric analysis*. Pearson.

Gu, B., Konana, P., Raghunathan, R., & Chen, H. M. (2014). Research note—The allure of homophily in social media: Evidence from investor responses on virtual communities. *Information Systems Research*, *25*(3), 604-617. https://doi.org/10.1287/isre.2014.0531

Gu, B., Konana, P., Rajagopalan, B., & Chen, H.-W. M. (2007). Competition among virtual communities and user valuation: The case of investing-related communities. *Information Systems Research*, *18*(1), 68-85. https://doi.org/10.1287/isre.1070.0114

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*. https://doi.org/10.1145/2872427.2883037

Huang, K.-Y., Chengalur-Smith, I., & Pinsonneault, A. (2019). Sharing is caring: Social support provision and companionship activities in healthcare virtual support communities. *MIS Quarterly*, *43*(2), 395-424. https://doi.org/10.25300/MISQ/2019/13225

Jabr, W., Mookerjee, R., Tan, Y., & Mookerjee, V. S. (2014). Leveraging philanthropic behavior for customer support: The case of user support forums. *MIS Quarterly*, *38*(1), 187-208. https://doi.org/10.25300/MISQ/2014/38.1.09

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, *62*(1), 85-96. https://doi.org/10.1111/j.1541-0420.2005.00396.x

Liu, X., Zhang, B., Susarlia, A., & Padman, R. (2020). Go to YouTube and call me in the morning: Use of social media for chronic conditions. *MIS Quarterly*, *44*(1), 257-283. https://doi.org/10.25300/MISQ/2020/15107

Lora, M. I., & Singer, J. M. (2008). Beta-binomial/Poisson regression models for repeated bivariate counts. *Statistics in Medicine*, *27*(17), 3366-3381. https://doi.org/10.1002/sim.3303

Lora, M. I., & Singer, J. M. (2011). Beta-binomial/gamma-Poisson regression models for repeated counts with random parameters. *Brazilian Journal of Probability and Statistics*, *25*(2), 218-235. https://doi.org/10.1214/10-BJPS118

Luo, X., Zhang, J. J., Gu, B., & Phang, C. (2013). Expert blogs and consumer perceptions of competing brands. *MIS Quarterly*, *41*(2), 371-395. https://doi.org/10.25300/MISQ/2017/41.2.03

Martin, B. D., Witten, D., & Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *Annals of Applied Statistics*, *14*(1), 94-115. https://doi.org/10.1214/19-AOAS1283

Moreno, A., & Terwiesch, C. (2014). Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research*, *25*(4), 865-886. https://doi.org/10.1214/19-AOAS1283

Murphy, K. M., & Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, *20*(1), 88-97. https://doi.org/10.1198/073500102753410417

Oberhofer, H., & Pfaffermayr, M. (2014). Two-part models for fractional responses defined as ratios of integers. *Econometrics*, *2*(3), 123-144. https://doi.org/10.3390/econometrics2030123

Qiao, M., & Huang, K.-W. (2021). Correcting misclassification bias in regression models with variables generated via data mining. *Information Systems Research*, *32*(2), 462-480. https://doi.org/10.1287/isre.2020.0977

Rumsey, D. J. (2006). *Probability for dummies*. John Wiley & Sons.

Schader, M., & Schmid, F. (1989). Two rules of thumb for the approximation of the binomial distribution by the normal distribution. *The American Statistician*, *43*(1), 23-24. https://doi.org/10.2307/2685162

Singh, P. V., Sahoo, N., & Mukhopadhyay, T. (2014). How to attract and retain readers in enterprise blogging? *Information Systems Research*, *25*(1), 35-52. https://doi.org/10.1287/isre.2013.0509

Wu, J., Huang, L., & Zhao, J. L. (2019). Operationalizing regulatory focus in the digital age: Evidence from an e-commerce context. *MIS Quarterly*, *43*(3), 745-764. https://doi.org/10.25300/MISQ/2019/14420

Yang, M., Adomavicius, G., Burtch, G., & Ren, Y. (2018). Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, *29*(1), 4-24. https://doi.org/10.1287/isre.2017.0727

Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2016). How much is an image worth? An empirical analysis of property's image aesthetic quality on demand at AirBNB. In *Proceedings of the International Conference on Information Systems*.

Zhu, J., Eickhoff, J. C., & Kaiser, M. S. (2003). Modeling the dependence between number of trials and success probability in beta-binomial-Poisson mixture distributions. *Biometrics*, *59*(4), 955-961. https://doi.org/10.1111/j.0006-341X.2003.00110.x

About the Authors

Mengke Qiao is an assistant professor of information systems at the Culverhouse College of Business, University of Alabama. She received her Ph.D. in information systems and analytics from the National University of Singapore. Her research interests focus on machine learning and causal inference. Her research has been published in *Information Systems Research*.

Ke-Wei Huang is the executive director of Asian Institute of Digital Finance and an associate professor in the Department of Information Systems and Analytics at the National University of Singapore (NUS). He received his Ph.D., M.Phil., and M.Sc. degrees in information systems from the Stern School of Business at New York University. His research interests include machine learning and causal inference, applied data mining in finance, IT labor economics and entrepreneurship, and the economics of IS (pricing digital goods). His research has been published in scholarly journals including *Information Systems Research*, *Strategic Management Journal*, and *Production and Operations Management*.

Appendix A

Technical Details of Case 1

Proof of Theorem 1

Given Assumption 2, we can derive the following two equations:

$$\mathrm{P}(X_i^g | W_i^1, \dots, W_i^{M_i}, Z_i, \bar{W}_i) \; = \mathrm{P}(X_i^g | W_i^g, Z_i, \bar{W}_i), \qquad (21)$$

$$\mathrm{P}(X_i^g | W_i^1, \dots, W_i^{M_i}, Z_i, \bar{W}_i, X_i^h) = \mathrm{P}(X_i^g | W_i^1, \dots, W_i^{M_i}, Z_i, \bar{W}_i). \qquad (22)$$

The first equation is derived based on the conditional independence between X_i^g and W_i^h. The second equation is based on the conditional independence between X_i^g and X_i^h.

Given these two equations, we can derive:

$$\begin{aligned} &\mathrm{P}(X_i^1, \dots, X_i^{M_i} | W_i^1, \dots, W_i^{M_i}, Z_i, \bar{W}_i) \\ &= \mathrm{P}(X_i^1 | W_i^1, \dots, W_i^{M_i}, Z_i, \bar{W}_i) \times \dots \times \mathrm{P}(X_i^{M_i} | W_i^1, \dots, W_i^{M_i}, Z_i, \bar{W}_i) \\ &= \mathrm{P}(X_i^1 | W_i^1, Z_i, \bar{W}_i) \times \dots \times \mathrm{P}(X_i^{M_i} | W_i^{M_i}, Z_i, \bar{W}_i) \; = \prod_{j=1}^{M_i} \mathrm{P}(X_i^j | W_i^j, Z_i, \bar{W}_i). \end{aligned}$$

The first equality is derived based on Equation (22) and the second equality is based on Equation (21).

Since there are only four values for $\mathrm{P}(X_i^j | W_i^j, Z_i, \bar{W}_i)$, given the combination of $X_i^1, \dots, X_i^{M_i}$ and $W_i^1, \dots, W_i^{M_i}$ with TP value as h, we can further derive:

$$\mathrm{P}(X_i^1, \dots, X_i^{M_i} | W_i^1, \dots, W_i^{M_i}, Z_i, \bar{W}_i) = \prod_{j=1}^{M_i} \mathrm{P}(X_i^j | W_i^j, Z_i, \bar{W}_i)$$

$$= \underbrace{(\mathrm{Pr}_i^{\mathrm{TP}})^h}_{A} \underbrace{(\mathrm{Pr}_i^{\mathrm{FP}})^{M_i\bar{W}_i - h}}_{B} \times \underbrace{(\mathrm{Pr}_i^{\mathrm{FN}})^{M_i\bar{X}_i - h}}_{C} \underbrace{(\mathrm{Pr}_i^{\mathrm{TN}})^{M_i - M_i\bar{W}_i - (M_i\bar{X}_i - h)}}_{D},$$

where terms A, B, C, and D represent the probabilities of the true positive, false positive, false negative, and true negative observations. There are several important observations from this expression. First, this probability depends only on $\bar{W}_i$ and M_i, but not the combination of $W_i^1, \dots, W_i^{M_i}$. Second, this probability also does not depend on the specific combination of $X_i^1, \dots, X_i^{M_i}$ while it only depends on $h, \bar{X}_i$, and M_i. In other words, the probabilities of all the combinations of $X_i^1, \dots, X_i^{M_i}$and $W_i^1, \dots, W_i^{M_i}$ with the same $\bar{X}_i$, $\bar{W}_i$, M_i, and h are the same. As a result, we can derive the conditional probability for each pair of $(\bar{X}_i, h)$, which is a multiplicative term of two binomial distribution probabilities:

$$\begin{aligned} \mathrm{P}(\bar{X}_i, h | \bar{W}_i, Z_i) = &\binom{M_i\bar{W}_i}{h} (\mathrm{Pr}_i^{\mathrm{TP}})^h (\mathrm{Pr}_i^{\mathrm{FP}})^{M_i\bar{W}_i - h} \times \binom{M_i - M_i\bar{W}_i}{M_i\bar{X}_i - h} (\mathrm{Pr}_i^{\mathrm{FN}})^{M_i\bar{X}_i - h} (\mathrm{Pr}_i^{\mathrm{TN}})^{M_i - M_i\bar{W}_i - (M_i\bar{X}_i - h)} \\ = &\; \mathrm{B}\big(h; M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{TP}}\big) \times \mathrm{B}\big(M_i\bar{X}_i - h; M_i - M_i\bar{W}_i, \mathrm{Pr}_i^{\mathrm{FN}}\big), \end{aligned}$$

where $\binom{M_i\bar{W}_i}{h}$ refers to the number of combinations where h true positive observations exist out of all the predicted positive observations $(M_i\bar{W}_i)$; $\binom{M_i - M_i\bar{W}_i}{M_i\bar{X}_i - h}$ refers to the number of combinations where $M_i\bar{X}_i - h$ false negative observations exist out of all the predicted negative observations $(M_i - M_i\bar{W}_i)$. Finally, we can further decompose $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i)$ by all possible values of true positive observations (captured by h), $\mathrm{P}(\bar{X}_i|\bar{W}_i, Z_i) = \sum_{h=0}^{M_i\bar{X}_i} \mathrm{P}(\bar{X}_i, h|\bar{W}_i, Z_i)$.

Example 1: Assume $\bar{X}_1$ is the average value of X (True review sentiment) from two reviews. The possible value of $\bar{X}_1$ is 0, 0.5, or 1. Similarly, the mean of the proxy variable $\bar{W}_1$ can also be 0, 0.5, or 1. Take $\bar{W}_1 = 0.5$ as an example, which indicates that one review is predicted as positive while the other is predicted as negative. Next, we can infer the conditional probability that $\bar{X}_1$ is 0, 0.5, or 1. If $\bar{X}_1 = 1$, then we have one TP review and one FN review. If $\bar{X}_1 = 0$, then we have one FP review and one TN review. The confusion matrices of these two examples are illustrated in Table A1 and Table A2. Given these two confusion matrices, we can calculate the conditional probability by Equation (6).

Table A1. Confusion Matrix ($\bar{X}_1 = 1$)

	$X_i^j = 1$	$X_i^j = 0$	Sum
$W_i^j = 1$	1	0	1
$W_i^j = 0$	1	0	1
Sum	2	0	2

Table A2. Confusion Matrix ($\bar{X}_1 = 0$)

	$X_i^j = 1$	$X_i^j = 0$	Sum
$W_i^j = 1$	0	1	1
$W_i^j = 0$	0	1	1
Sum	0	2	2

Table A3. Confusion Matrix 1 ($h = 0$)

	$X_i^j = 1$	$X_i^j = 0$	Sum
$W_i^j = 1$	0	1	1
$W_i^j = 0$	1	0	1
Sum	1	1	2

Table A4. Confusion Matrix 2 ($h = 1$)

	$X_i^j = 1$	$X_i^j = 0$	Sum
$W_i^j = 1$	1	0	1
$W_i^j = 0$	0	1	1
Sum	1	1	2

If $\bar{X}_1 = 0.5$, there are two scenarios and it is the more complicated case. Either both reviews are predicted correctly or predicted wrongly. This is the additional loop over h in Equation (7). We can derive two confusion matrices, which are illustrated in Table A3 and Table A4.These two matrices correspond to two possible numbers of true positive reviews (h). When h equals 0 or 1, we can derive the first or second confusion matrix. Mathematically, $\mathrm{P}(\bar{X}_1 = 0.5|\bar{W}_1, Z_1)$ can be expressed as:

$$\mathrm{P}(\bar{X}_1 = 0.5|\bar{W}_1, Z_1) = \mathrm{P}(\bar{X}_1 = 0.5, h = 0|\bar{W}_1, Z_1) + \mathrm{P}(\bar{X}_1 = 0.5, h = 1|\bar{W}_1, Z_1),$$

$$\text{where } \mathrm{P}(\bar{X}_1 = 0.5, h = 0|\bar{W}_1, Z_1) = \mathrm{B}(0; 1, \mathrm{Pr}^{\mathrm{TP}})\times \mathrm{B}(1; 1, \mathrm{Pr}^{\mathrm{FN}}) \text{ and}$$

$$\mathrm{P}(\bar{X}_1 = 0.5, h = 1|\bar{W}_1, Z_1) = \mathrm{B}(1; 1, \mathrm{Pr}^{\mathrm{TP}})\times \mathrm{B}(0; 1, \mathrm{Pr}^{\mathrm{FN}}).$$

Proof of Approximated Solution by Normal Distribution

For predicted positive observations, the number of TP observations follows a binomial distribution, $\mathrm{B}(\cdot\,; \mathrm{WP}, \mathrm{Pr}_i^{\mathrm{TP}})$. The mean and variance of this binomial distribution are,

$$\mathrm{E}(\mathrm{TP}|\bar{W}_i, Z_i) = \mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}}, \ \ \mathrm{Var}(\mathrm{TP}|\bar{W}_i, Z_i) = \mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}} \times \mathrm{Pr}_i^{\mathrm{FP}}.$$

For predicted negative observations, the number of FN observations follows a binomial distribution, $\mathrm{B}(\cdot\,; \mathrm{WN}, \mathrm{Pr}_i^{\mathrm{FN}})$. The mean and variance of this binomial distribution are:

$$\mathrm{E}(\mathrm{FN}|\bar{W}_i, Z_i) = \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}}, \ \ \mathrm{Var}(\mathrm{FN}|\bar{W}_i, Z_i) = \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}} \times \mathrm{Pr}_i^{\mathrm{TN}}.$$

When the sample size is large enough, two binomial distributions can be approximated by the normal distribution. Since $\bar{X}_i$ is a linear transformation of TP and FN ($\frac{\mathrm{TP+FN}}{M_i}$), we can derive the mean and variance of the normal distribution of $\bar{X}_i$ by the following formulas:

$$\mu_i = \mathrm{E}(\bar{X}_i|\bar{W}_i, Z_i) = \frac{1}{M_i}[\mathrm{E}(\mathrm{TP}|\bar{W}_i, Z_i) + \mathrm{E}(\mathrm{FN}|\bar{W}_i, Z_i)] = \frac{1}{M_i}\left[\mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}} + \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}}\right].$$

$$\begin{aligned}\sigma_i^2 = \mathrm{Var}(\bar{X}_i|\bar{W}_i, Z_i) &= \frac{1}{{M_i}^2}[\mathrm{Var}(\mathrm{TP}|\bar{W}_i, Z_i) + \mathrm{Var}(\mathrm{FN}|\bar{W}_i, Z_i) + \mathrm{Cov}(\mathrm{TP}, \mathrm{FN}|\bar{W}_i, Z_i)] \\ &= \frac{1}{{M_i}^2}\left[\mathrm{WP} \times \mathrm{Pr}_i^{\mathrm{TP}} \times \mathrm{Pr}_i^{\mathrm{FP}} \ + \mathrm{WN} \times \mathrm{Pr}_i^{\mathrm{FN}} \times \mathrm{Pr}_i^{\mathrm{TN}}\right],\end{aligned}$$

where $\mathrm{Cov}(\mathrm{TP}, \mathrm{FN}|\bar{W}_i, Z_i) = 0$ since X_i^j of the observations are conditionally independent under Assumption 2.

Extension to Multiple Focal Independent Variables

Assume we have two focal independent variables ($\bar{X}_{1i}$ and $\bar{X}_{2i}$) constructed from two individual-level variables,[13] X_{1i}^j and X_{2i}^j. For example, we have 1000 products and each product has some reviews. Each review has two proxy variables, one is review sentiment with two labels (positive or negative), and the other is review subjectivity with two labels (objective or subjective). $\bar{X}_{1i}$ is the mean of sentiment labels and $\bar{X}_{2i}$ is the mean of the subjectivity labels for product i. The true regression model is defined as $\mathrm{P}(Y_i|\bar{X}_{1i}, \bar{X}_{2i}, Z_i)$. Then, following the logic of Equation (5), our corrected probability function is:

$$\mathrm{P}(Y_i|\bar{W}_{1i}, \bar{W}_{2i}, Z_i) = \sum_{\bar{X}_{1i}, \bar{X}_{2i}} \mathrm{P}(Y_i|\bar{X}_{1i}, \bar{X}_{2i}, Z_i)\mathrm{P}(\bar{X}_{1i}, \bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i).$$

where $\mathrm{P}(\bar{X}_{1i}, \bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i)$ is the aggregate-level measurement error function. Next, we decompose this function into two terms where each term has only one focal variable:

$$\begin{aligned}\mathrm{P}(\bar{X}_{1i}, \bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i) &= \mathrm{P}(\bar{X}_{1i}|\bar{X}_{2i}, \bar{W}_{1i}, \bar{W}_{2i}, Z_i) \times \mathrm{P}(\bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i) \\ &= \mathrm{P}(\bar{X}_{1i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i) \times \mathrm{P}(\bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i),\end{aligned}$$

where $\mathrm{P}(\bar{X}_{1i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i)$ and $\mathrm{P}(\bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i)$ can be estimated following the methods in the Theoretical Solution for Case 1 section. The assumption of this decomposition is that $\bar{X}_{1i}$ and $\bar{X}_{2i}$ are conditionally independent on $\bar{W}_{1i}, \bar{W}_{2i}$, and Z_i. This assumption implies that $\bar{X}_{2i}$ cannot provide additional information for inferring the probability function of $\bar{X}_{1i}$ conditional on $(\bar{W}_{1i}, \bar{W}_{2i}, Z_i)$. In other words, we assume $\mathrm{P}(\bar{X}_{1i}|\bar{X}_{2i}, \bar{W}_{1i}, \bar{W}_{2i}, Z_i) = \mathrm{P}(\bar{X}_{1i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i)$. The reason for imposing this assumption is to simplify the joint probability by the product of two marginal probabilities with only one focal variable, which can be estimated by Equation (6).

[13] The other case is two variables are constructed from one proxy variable with three classes. In this case, to derive $\mathrm{P}(\bar{X}_{1i}, \bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i)$, researchers need to construct a confusion matrix with three classes. The individual-level measurement error model is $p_{tk} = P\left(X_i^j = k\middle|W_i^j = t, \bar{W}_{1i}, \bar{W}_{2i}, Z_i\right), t = 1,2,3$ and $k = 1,2,3$. Finally, researchers can combine the confusion matrix and p_{tk} to derive $\mathrm{P}(\bar{X}_{1i}, \bar{X}_{2i}|\bar{W}_{1i}, \bar{W}_{2i}, Z_i)$ by following the logic in the proof of Theorem 1.

Appendix B

Technical Details of Case 3

Proof of Theorem 4

Given Assumption 3, we can derive the following two equations:

$$\mathrm{P}(y_i^g|Y_i^1, \dots, Y_i^{M_i}, X_i) = \mathrm{P}(y_i^g|Y_i^g, X_i),$$

$$\mathrm{P}(y_i^g|Y_i^1, \dots, Y_i^{M_i}, X_i, y_i^h) = \mathrm{P}(y_i^g|Y_i^1, \dots, Y_i^{M_i}, X_i).$$

The first equation is derived based on the conditional independence between y_i^g and Y_i^h. The second equation is based on the conditional independence between y_i^g and y_i^h.

Given these two equations, we can derive, $\mathrm{P}(y_i^1, \dots, y_i^{M_i}|Y_i^1, \dots, Y_i^{M_i}, X_i) = \prod_{j=1}^{M_i} \mathrm{P}(y_i^j|Y_i^j, X_i)$.

Since there are only four values for $\mathrm{P}(y_i^j|Y_i^j, X_i)$, given the combination of $y_i^1, \dots, y_i^{M_i}$ and $Y_i^1, \dots, Y_i^{M_i}$ with TP value as h, we can further derive:

$$\mathrm{P}(y_i^1, \dots, y_i^{M_i}|Y_i^1, \dots, Y_i^{M_i}, X_i) = \prod_{j=1}^{M_i} \mathrm{P}(y_i^j|Y_i^j, X_i)$$
$$= \underbrace{\left(\mathrm{Pr}_i^{\mathrm{TP}}\right)^h}_{A} \underbrace{\left(\mathrm{Pr}_i^{\mathrm{FN}}\right)^{M_i\bar{Y}_i - h}}_{B} \times \underbrace{\left(\mathrm{Pr}_i^{\mathrm{FP}}\right)^{M_i\bar{y}_i - h}}_{C} \underbrace{\left(\mathrm{Pr}_i^{\mathrm{TN}}\right)^{M_i - M_i\bar{Y}_i - (M_i\bar{y}_i - h)}}_{D}.$$

Next, we derive the conditional probability for each pair of $(\bar{y}_i, h)$, which is a multiplicative term of two binomial distribution probabilities:

$$\begin{aligned}\mathrm{P}(\bar{y}_i, h|\bar{Y}_i, X_i) = {} & \binom{M_i\bar{Y}_i}{h} \left(\mathrm{Pr}_i^{\mathrm{TP}}\right)^h \left(\mathrm{Pr}_i^{\mathrm{FN}}\right)^{M_i\bar{Y}_i - h} \times \binom{M_i - M_i\bar{Y}_i}{M_i\bar{y}_i - h} \left(\mathrm{Pr}_i^{\mathrm{FP}}\right)^{M_i\bar{y}_i - h} \left(\mathrm{Pr}_i^{\mathrm{TN}}\right)^{\substack{M_i - M_i\bar{Y}_i \\ -(M_i\bar{y}_i - h)}} \\ = {} & \mathrm{B}\left(h; M_i\bar{Y}_i, \mathrm{Pr}_i^{\mathrm{TP}}\right) \times \mathrm{B}\left(M_i\bar{y}_i - h; M_i - M_i\bar{Y}_i, \mathrm{Pr}_i^{\mathrm{FP}}\right).\end{aligned}$$

Finally, we can decompose $\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i)$ by all possible values of true positive observations, $\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i) = \sum_{h=0}^{M_i\bar{y}_i} \mathrm{P}(\bar{y}_i, h|\bar{Y}_i, X_i)$.

Measurement Error Model 2: Approximated Solution by Normal Distribution

To simplify the computation, we can apply Lyapunov central limit theorem to approximate the distribution of the mean of y_i^j by the normal distribution. By setting $M_i\bar{Y}_i$ as YP and $M_i - M_i\bar{Y}_i$ as YN, the mean and variance of the normal density function $(f(\bar{y}_i|\bar{Y}_i, X_i))$ are given by:

$$\mu_i = \mathrm{E}(\bar{y}_i|\bar{Y}_i, X_i) = \tfrac{1}{M_i}\left[\mathrm{YP} \times \mathrm{Pr}_i^{\mathrm{TP}} + \mathrm{YN} \times \mathrm{Pr}_i^{\mathrm{FP}}\right],$$

$$\sigma_i^2 = \mathrm{Var}(\bar{y}_i|\bar{Y}_i, X_i) = \frac{1}{{M_i}^2}\left[\mathrm{YP} \times \mathrm{Pr}_i^{\mathrm{TP}} \times \mathrm{Pr}_i^{\mathrm{FN}} + \mathrm{YN} \times \mathrm{Pr}_i^{\mathrm{FP}} \times \mathrm{Pr}_i^{\mathrm{TN}}\right],$$

where $\mathrm{Cov}(\mathrm{TP}, \mathrm{FP}|\bar{Y}_i, X_i) = 0$ since y_i^j of the observations are conditionally independent given Assumption 3. Next, we utilize half-unit continuity correction to derive $\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i)$. Let Φ denote the cumulative distribution function of $f(\bar{y}_i|\bar{Y}_i, X_i)$:

$$\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i) = \Phi\left(\bar{y}_i + \tfrac{1}{M_i} \times 0.5\right) - \Phi(\bar{y}_i - \tfrac{1}{M_i} \times 0.5). \qquad (23)$$

The rule of normal approximation by CLT is that all the values within three standard deviations around $\mathrm{E}(\bar{y}_i|\bar{Y}_i, X_i)$ fall within the range of $\bar{y}_i$ (0 to 1). All technical details of this part are qualitatively the same as Case 1.

Theorem 5: *Let the labeled dataset be the random sample i.i.d drawn from the population. Suppose Assumption 3 holds, β in Equation (11) can be approximately estimated by applying MLE to Equation (13) where $\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i)$ is estimated by Equation (23) when M_i is large enough.*

Measurement Error Model 3: Approximated Solution by Law of Large Numbers.

When M_i is large enough, $\bar{y}_i$ converges to its conditional expectation $\mathrm{E}(\bar{y}_i|\bar{Y}_i, X_i)$ by the strong law of large numbers (Feller, 2008):

$$\mathrm{P}(\bar{y}_i = \mathrm{E}(\bar{y}_i|\bar{Y}_i, X_i)|\bar{Y}_i, X_i) = 1. \qquad (24)$$

By modifying equality, $\bar{y}_i = \mathrm{E}(\bar{y}_i|\bar{Y}_i, X_i)$, we derive the formula of $\bar{Y}_i$ as follows:

$$\bar{Y}_i^{\ *} = \frac{\bar{y}_i - \mathrm{Pr}_i^{\mathrm{FP}}}{\mathrm{Pr}_i^{\mathrm{TP}} - \mathrm{Pr}_i^{\mathrm{FP}}}.$$

Finally, Equation (13) can be simplified to:

$$\mathrm{P}(\bar{y}_i|X_i) = \mathrm{P}\big(\bar{y}_i = \mathrm{E}\big(\bar{y}_i|\bar{Y}_i^{\ *}, X_i\big)|\bar{Y}_i^{\ *}, X_i\big)\mathrm{P}(\bar{Y}_i^{\ *}|X_i) = \mathrm{P}(\bar{Y}_i^{\ *}|X_i).$$

The right side of the first equality only has the term when $\bar{y}_i = \mathrm{E}\big(\bar{y}_i|\bar{Y}_i^{\ *}, X_i\big)$ since Equation (16) implies that $\mathrm{P}\big(\bar{y}_i|\bar{Y}_i^{\ *}, X_i\big) = 0$ for all other possible values of $\bar{Y}_i$.

Theorem 6: *Let the labeled dataset be the random sample i.i.d drawn from the population. Suppose Assumption 3 holds, β in Equation (11) can be consistently estimated by applying MLE to Equation (13) where $\mathrm{P}(\bar{y}_i|\bar{Y}_i, X_i)$ is estimated by Equation (24) when M_i is large enough.*

Appendix C

Proof of Theoretical Analysis of Estimation Bias

Technical Details of Expected Value and Variance of Aggregated Measurement Error

Let $\overline{W}_i - \bar{X}_i = \bar{X}_i \times \left(\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + (1 - \bar{X}_i) \times \widehat{\mathrm{Pr}}_i^{\mathrm{FP}}$. The expectation of this measurement error is derived from the law of total expectation:

$$\mathrm{E}(\overline{W}_i - \bar{X}_i) = \mathrm{E}\left(\mathrm{E}\left(\bar{X}_i \times \left(\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + (1 - \bar{X}_i) \times \widehat{\mathrm{Pr}}_i^{\mathrm{FP}} \middle| \bar{X}_i\right)\right)$$

$$= \mathrm{E}\left(\bar{X}_i \times \left(\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + (1 - \bar{X}_i) \times \overline{\mathrm{Pr}}_i^{\mathrm{FP}}\right) = \mathrm{E}(\bar{X}_i) \times \left(\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + \left(1 - \mathrm{E}(\bar{X}_i)\right) \times \overline{\mathrm{Pr}}_i^{\mathrm{FP}}.$$

The variance of measurement error is derived by the law of total variance:

$$\mathrm{Var}(\overline{W}_i - \bar{X}_i) = \mathrm{E}\left(\mathrm{Var}(\overline{W}_i - \bar{X}_i | \bar{X}_i)\right) + \mathrm{Var}(\mathrm{E}(\overline{W}_i - \bar{X}_i | \bar{X}_i))$$

$$= \mathrm{E}\left(\frac{\bar{X}_i \overline{\mathrm{Pr}}_i^{\mathrm{TP}}\left(1 - \overline{\mathrm{Pr}}_i^{\mathrm{TP}}\right) + (1 - \bar{X}_i)\overline{\mathrm{Pr}}_i^{\mathrm{FP}}\left(1 - \overline{\mathrm{Pr}}_i^{\mathrm{FP}}\right)}{M_i}\right) + \mathrm{Var}(\bar{X}_i \times \left(\overline{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + (1 - \bar{X}_i) \times \overline{\mathrm{Pr}}_i^{\mathrm{FP}})$$

$$= \frac{\mathrm{E}(\bar{X}_i)\overline{\mathrm{Pr}}_i^{\mathrm{TP}}\left(1 - \overline{\mathrm{Pr}}_i^{\mathrm{TP}}\right) + \left(1 - \mathrm{E}(\bar{X}_i)\right)\overline{\mathrm{Pr}}_i^{\mathrm{FP}}\left(1 - \overline{\mathrm{Pr}}_i^{\mathrm{FP}}\right)}{M_i} + \left(\overline{\mathrm{Pr}}_i^{\mathrm{FN}} + \overline{\mathrm{Pr}}_i^{\mathrm{FP}}\right)^2 \mathrm{Var}(\bar{X}_i).$$

Technical Details of Mean of Proxy Variable as the Focal Independent Variable

Let $\overline{W}_i - \bar{X}_i = \bar{e}_i = \mathrm{E}(\bar{e}_i) + \mu_i$. Given $\mathrm{E}(\bar{X}_i) = \left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}}\right)\mathrm{E}(\overline{W}_i) - \frac{\overline{\mathrm{Pr}}^{\mathrm{FP}}}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}}$, we derive $\mathrm{E}(\bar{e}_i)$ as $-\left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} - 1\right)\mathrm{E}(\overline{W}_i) + \frac{\overline{\mathrm{Pr}}^{\mathrm{FP}}}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}}$.

$$\lim_{N \to \infty} \hat{\beta} = \beta + \beta \frac{\mathrm{Cov}(-\bar{e}_i, \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)} = \beta + \beta \frac{\mathrm{Cov}(-\mathrm{E}(\bar{e}_i), \overline{W}_i)}{Var(\overline{W}_i)} + \beta \frac{\mathrm{Cov}(-\mu_i, \overline{W}_i)}{Var(\overline{W}_i)}$$

$$= \beta + \beta\left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} - 1\right)\frac{\mathrm{Cov}(\mathrm{E}(\overline{W}_i), \overline{W}_i)}{\mathrm{Var}(\overline{W}_i)} + \beta \frac{\mathrm{Cov}(-\mu_i, \overline{W}_i)}{Var(\overline{W}_i)}.$$

Technical Details of Mean of Proxy Variable as the Dependent Variable

Let $\bar{e}_i = \mathrm{E}(\bar{e}_i) + \mu_i$, where $\mathrm{E}(\bar{e}_i) = X_i\beta \times (\overline{\mathrm{Pr}}^{\mathrm{TP}} - 1) + (1 - X_i\beta) \times \overline{\mathrm{Pr}}^{\mathrm{FP}}$conditional on X_i by the law of total expectation. We derive the formula of estimated β by replacing e_i by $\bar{e}_i$ in Equation (2):

$$\lim_{N \to \infty} \hat{\beta} = \beta + \frac{\mathrm{Cov}(\bar{e}_i, X_i)}{\mathrm{Var}(X_i)} = \beta + \frac{\mathrm{Cov}(\mathrm{E}(\bar{e}_i), X_i)}{\mathrm{Var}(X_i)} + \frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)}$$

$$= \beta + (\overline{\mathrm{Pr}}^{\mathrm{TP}} - 1 - \overline{\mathrm{Pr}}^{\mathrm{FP}})\beta + \frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)} = (\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}})\beta,$$

where $\frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)} = 0$. The proof for $\frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)}$ equaling 0 is based on covariance formula $\mathrm{Cov}(A, B) = \mathrm{E}(AB) - \mathrm{E}(A)\mathrm{E}(B)$ and the law of total expectation:

$$\mu_i = (X_i\beta) \times \left[\left(\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{TP}}\right) - \left(\widehat{\mathrm{Pr}}_i^{\mathrm{FP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}\right)\right] + \widehat{\mathrm{Pr}}_i^{\mathrm{FP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}} + (\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - \widehat{\mathrm{Pr}}_i^{\mathrm{FP}} - 1)\varepsilon_i.$$

$$\mathrm{Cov}(\mu_i, X_i) = \mathrm{E}(\mu_i X_i) - \mathrm{E}(\mu_i)\mathrm{E}(X_i) = 0.$$

Due to the page limit, we omitted the details.

Technical Details for Sum of Proxy Variable as the Focal Independent Variable

Let $S_{W_i} - S_{X_i} = S_{e_i} = \mathrm{E}\left(S_{e_i}\right) + \mu_i$, where $\mathrm{E}\left(S_{e_i}\right) =- \left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} - 1\right) \mathrm{E}\left(S_{W_i}\right) + \frac{\overline{\mathrm{Pr}}^{\mathrm{FP}}}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} M_i$. Then, we derive,

$$\lim_{N\to\infty} \hat{\beta} = \beta + \beta \left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} - 1\right) \frac{\mathrm{Cov}\left(\mathrm{E}\left(S_{W_i}\right), S_{W_i}\right)}{\mathrm{Var}\left(S_{W_i}\right)} - \beta \frac{\overline{\mathrm{Pr}}^{\mathrm{FP}}}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} \rho + \beta \frac{\mathrm{Cov}\left(\mu_i, S_{W_i}\right)}{\mathrm{Var}\left(S_{W_i}\right)},$$

where $\rho = \frac{\mathrm{Cov}(M_i, S_{W_i})}{\mathrm{Var}(S_{W_i})}$. Next, we define $M_i{}^*$ as the maximum value of M_i, and $s_i = \frac{M_i}{M_i{}^*}$.

$$\frac{\mathrm{Cov}\left(\mathrm{E}\left(S_{W_i}\right), S_{W_i}\right)}{\mathrm{Var}\left(S_{W_i}\right)} = \frac{\mathrm{Cov}(s_i \mathrm{E}(\overline{W}_i), s_i \overline{W}_i)}{\mathrm{Var}(s_i \overline{W}_i)}, \qquad \frac{\mathrm{Cov}\left(\mu_i, S_{W_i}\right)}{\mathrm{Var}\left(S_{W_i}\right)} = \frac{\mathrm{Cov}(s_i \mu_i / M_i, s_i \overline{W}_i)}{\mathrm{Var}(s_i \overline{W}_i)}.$$

Since s_i is a ratio between 0 and 1, this transformation enables us to apply the law of large numbers to simplify the formula. Specifically, when M_i is large enough, $\overline{W}_i$ will converge to $\mathrm{E}(\overline{W}_i)$ and μ_i / M_i will converge to 0 by the law of large numbers. Therefore, $\frac{\mathrm{Cov}(s_i \mathrm{E}(\overline{W}_i), s_i \overline{W}_i)}{\mathrm{Var}(s_i \overline{W}_i)} = 1$ and $\frac{\mathrm{Cov}(s_i \mu_i / M_i, s_i \overline{W}_i)}{\mathrm{Var}(s_i \overline{W}_i)} = 0$. As a result:

$$\lim_{N\to\infty, M_i \to\infty} \hat{\beta} = \beta \left(\frac{1}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}}\right) - \beta \frac{\overline{\mathrm{Pr}}^{\mathrm{FP}}}{\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}} \rho.$$

Technical Details for Sum of Proxy Variable as the Dependent Variable

Let $S_{e_i} = S_{y_i} - S_{Y_i} = S_{Y_i} \times \left(\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - 1\right) + \left(M_i - S_{Y_i}\right) \times \widehat{\mathrm{Pr}}_i^{\mathrm{FP}}$. Further, $S_{e_i} = \mathrm{E}\left(S_{e_i}\right) + \mu_i$, where $\mathrm{E}\left(S_{e_i}\right) = (X_i \beta \times (\overline{\mathrm{Pr}}^{\mathrm{TP}} - 1) + (M_i - X_i \beta) \times \overline{\mathrm{Pr}}^{\mathrm{FP}})$ conditional on X_i by the law of total expectation, and $\mu_i = (X_i \beta) \times \left[\left(\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{TP}}\right) - \left(\widehat{\mathrm{Pr}}_i^{\mathrm{FP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}\right)\right] + M_i \times (\widehat{\mathrm{Pr}}_i^{\mathrm{FP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}) + (\widehat{\mathrm{Pr}}_i^{\mathrm{TP}} - \widehat{\mathrm{Pr}}_i^{\mathrm{FP}} - 1) \varepsilon_i$. Next, we obtain:

$$\lim_{N\to\infty} \hat{\beta} = \beta + \frac{\mathrm{Cov}(S_{e_i}, X_i)}{\mathrm{Var}(X_i)} = \beta + \frac{\mathrm{Cov}\left(\mathrm{E}\left(S_{e_i}\right), X_i\right)}{\mathrm{Var}(X_i)} + \frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)}$$

$$= \beta + \beta(\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}} - 1) + \mathrm{Pr}^{\mathrm{FP}} \varphi + \frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)} = \beta(\overline{\mathrm{Pr}}^{\mathrm{TP}} - \overline{\mathrm{Pr}}^{\mathrm{FP}}) + \overline{\mathrm{Pr}}^{\mathrm{FP}} \varphi,$$

where $\varphi = \frac{\mathrm{Cov}(M_i, X_i)}{\mathrm{Var}(X_i)}$ and $\frac{\mathrm{Cov}(\mu_i, X_i)}{\mathrm{Var}(X_i)} = 0$. Following a similar logic to the third part of this Appendix, we can prove that $\mathrm{Cov}(\mu_i, X_i) = 0$.