

---

# Methods

---

## Power of Tests for a Dichotomous Independent Variable Measured with Error

*Daniel F. McCaffrey and Marc N. Elliott*

---

**Objective.** To examine the implications for statistical power of using predicted probabilities for a dichotomous independent variable, rather than the actual variable.

**Data Sources/Study Setting.** An application uses 271,479 observations from the 2000 to 2002 CAHPS Medicare Fee-for-Service surveys.

**Study Design and Data.** A methodological study with simulation results and a substantive application to previously collected data.

**Principle Findings.** Researchers often must employ key dichotomous predictors that are unobserved but for which predictions exist. We consider three approaches to such data: the *classification estimator* (1); the *direct substitution estimator* (2); the *partial information maximum likelihood estimator* (3, PIMLE). The efficiency of (1) (its power relative to testing with the true variable) roughly scales with the square of one less the classification error. The efficiency of (2) roughly scales with the  $R^2$  for predicting the unobserved dichotomous variable, and is usually more powerful than (1). Approach (3) is most powerful, but for testing differences in means of 0.2–0.5 standard deviations, (2) is typically more than 95 percent as efficient as (3).

**Conclusions.** The information loss from not observing actual values of dichotomous predictors can be quite large. Direct substitution is easy to implement and interpret and nearly as efficient as the PIMLE.

**Key Words.** Geocoding, measurement error, categorical data, methodology

---

Health services researchers often face measurement errors in the independent variables they wish to use in regression models. In particular, analysts might not observe a dichotomous (or categorical) variable identifying group membership for every observation, but might have estimates of  $p$ , the probability of being in the group from a rich model estimated on auxiliary data. This situation arises in many settings. For example, in the illustration presented later in this paper, researchers developed a model for predicting depression that can be used to measure the association between depression and

evaluations of care for beneficiaries in other Medicare databases. Diagnostic status as measured by a short screening instrument or demographic characteristics measured by geocoding from Census block groups are other examples of error prone categorical variables used in health services research.

Current approaches typically classify probabilistic information about category membership and then analyze the data as if these characteristics were measured without error. However, several approaches exist for testing hypotheses about the coefficient of the unobserved dichotomous predictor in linear models for an outcome of interest. Three estimators summarize the range of possible approaches. These are (1) the *classification estimator*, which classifies observations as zero or one on the basis of whether the probability that the unobserved variable is one exceeds a threshold; (2) the *direct substitution estimator*, which uses the observed probability  $p$  in place of the unobserved predictor in a regression of the outcome on  $p$ ; and (3) the *partial information maximum likelihood estimator (PIMLE)* which is the maximum likelihood estimator given that we only know  $p$ . We will compare these three estimators to a reference estimator that is only possible when the dichotomous predictor is known, which we call the *full information estimator*.

The classification and direct substitution estimators can be implemented using standard linear regression software without additional coding, and are the most widely used methods in practice. Classification has the additional appeal of providing intuitive groupings of all observations, so that simple summaries can be used in conjunction with modeling. Modeling with the probabilities is easy to implement, but less appealing to some analysts because the model is fit with a continuous variable and because the data are not categorized. Although optimization algorithms are available in common statistical software packages, maximum likelihood estimation, such as that used with the PIMLE, requires greater quantitative expertise to implement than the other methods. As a result, it is less accessible to many analysts who nonetheless routinely employ estimated dichotomous predictors.

Classification by the probability of an unobserved dichotomous predictor results in measurement error in the independent variables of the analysis for which there is an extensive literature (Fuller 1987 provides a good introduction to this literature). The three approaches considered here are each analogous to particular methods in the broader measurement error literature;

---

Address correspondence to Daniel F. McCaffrey, Ph.D., RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213. Marc N. Elliott, Ph.D., is with RAND Corporation, 1776 Main Street, Santa Monica, CA.

here we discuss these approaches in a unified context. The classification estimator has the same properties as regression with error-prone dichotomous independent variables. The direct substitution estimator is analogous to a two-stage instrumental variables (IV) estimator, where the probabilities from the external source in our problem are like the probabilities estimated using the IV in the first stage of the IV estimator (Fuller 1987). The PIMLE is analogous to using an IV via maximum likelihood estimation (Fuller 1987), except that in our problem we observed the probabilities rather than the IV. The PIMLE is also analogous to imputing the unobserved dichotomous variable using multiple imputation (Rubin 1987).

In general the literature finds that measurement error can result in loss of power for testing hypotheses; Tevere, Sobel, and Gilles (1995) specifically discuss the loss in power from modeling with an error-prone dichotomous predictor. However, a review of this literature did not find direct guidance for choosing among the straightforward and somewhat more complex approaches (e.g., PIMLE) in terms of the power of each approach.

This paper specifically addresses the power for testing hypotheses using the alternative approaches for the problem of unobserved dichotomous predictor values. We provide analytic formulas for the loss in power from using classification or direct substitution relative to the power from the full information estimator. Through a simulation study, we compare the power of these methods to those of the PIMLE. As the literature predicts, the loss of power from measurement error can be substantial and PIMLE provides more power than the alternatives. Somewhat surprisingly, the power of the direct substitution estimator and the PIMLE are quite similar when the true effect size of an unobserved dichotomous predictor is small to moderate.

The next section presents details of the estimators and analytic results. The following section describes a simulation study. The last section provides an empirical illustration of these results, and the paper concludes with a discussion of the implications of our findings.

## POWER OF CLASSIFICATION, DIRECT SUBSTITUTION, AND PARTIAL INFORMATION MLE

We consider the following simple linear model for a continuous dependent variable  $Y_i$ :

$$Y_i = \alpha + \beta Z_i + \varepsilon_i \quad (1)$$

for  $i = 1, \dots, n$  where  $Z_i$  is a dichotomous predictor that is independent of  $\varepsilon_i$ . The  $\varepsilon_i$  have mean zero and variance  $\sigma^2$  and  $\Pr(Z_i|X_i, \varepsilon_i) = \Pr(Z_i|X_i)$  for a given set of variables,  $X$ , used to predict the dichotomous predictor. We assume that  $Z_i$  is unobserved but that  $p_i = \Pr(Z_i = 1|X_i)$  is observed. We will primarily treat  $p_i$  as known without estimation error. We assume that the average of the  $p_i$  equals  $p$ , the overall mean of  $Z_i$ . The goal of the study is to make inferences about  $\beta$ , the difference between the means of the groups defined by  $Z$ . In particular, we will consider testing the null hypothesis,  $H_0: \beta = 0$ .

Our interest is in determining the power of tests based on the methods presented above. The power of the test depends on the parameter  $T$ , which is equal to the expected value of the estimator of  $\beta$  divided by its standard error (Steel and Torrie 1980). Thus for each method, we will determine  $T$  and compare it with  $T_0$ , the parameter for a test based on modeling with the unobservable  $Z$ s. To make this comparison more readily interpretable, we consider the square of the ratio of  $T$  to  $T_0$ , which we call the efficiency of the method. The reciprocal of the efficiency equals the ratio of the sample sizes required to achieve equal power using one of the methods based on the predicted values and an analysis using the unobservable  $Z$ s.

### *Classification Estimator*

*Properties.* The classification estimator of  $\beta$  when  $Z$  is unknown classifies  $p_i$  to create dichotomous predictors,  $U_i = 1$  when  $p_i$  is greater than threshold  $p^*$ , and 0 otherwise. Testing employs a standard independent sample  $t$ -test, with  $U_i$  in place of the unobserved  $Z_i$ .

One challenge for this method is how to select  $p^*$ ; a variety of procedures are commonly employed to choose this value. Selecting an approach is not straightforward, because the rule that yields the most power depends on the distribution of  $p_i$ , which itself is likely to be a function of both its true mean and the method by which it was estimated. We briefly discuss four common cutoffs for dichotomous classification. Simulations not presented here<sup>1</sup> suggest that none of these cutoffs is uniformly superior to the others, but we focus on one that generally performed the best.

Perhaps the most common approach is to split at  $p^* = .5$ . This approach fares very poorly when  $p$  is near 0 or 1, as it can result in massive misclassification. A second approach, meant to minimize variance, sets  $p^*$  at the median of the  $p_i$ . A third approach assumes bimodality in the

$p_i$  corresponding to true values of  $Z_i$ , and so sets  $p^*$  at the percentile of the  $p_i$  that corresponds to the mean of the  $p_i$ . The fourth and generally best of these approaches sets  $p^*$  at the mean of the  $p_i$ . This approach takes advantage of the  $p_i$  as unbiased estimators of  $p$ .

As detailed in Section A.1 of the supplementary appendix, the classification estimator results in an attenuated estimate of the effect of the independent variable. The expected value of the classification estimator is the product of the true effect size and an attenuation factor equal to the difference in the proportion of true “1s” among those classified as “1” and those classified as “0.” The classification estimator may have more or less variance than the full information estimator, because (a) the variance of the estimator is closely tied to the proportion of cases treated as “1,” and (b) the proportion classified as “1” may differ between classification and the full-information approach.

The efficiency from classification on  $U$  rather than  $Z$  (assuming the variance is known) is

$$\text{Classification efficiency} = \frac{(r_1 - r_0)^2 a(1 - a)}{\{[r_0(1 - r_0)a + r_1(1 - r_1)(1 - a)]E^2 + 1\}p(1 - p)} \tag{2}$$

where  $E = \beta/\sigma$  is the true effect size,  $a$  is the proportion of observations classified as “1,”  $r_1 - r_0$  is the attenuation factor with  $r_1$  equal to the proportion of cases classified as “1” that are true “1s,” and  $r_0$  equal to the proportion of cases classified as “0” that are true “1s.” The appendix shows that the power from classification equals that from the full information estimator only under perfect classification; otherwise classification efficiency is  $< 1$ .

Attenuation is the major source of loss of power. Correcting for attenuation by dividing estimates by the amount of attenuation merely exchanges bias for variance and does not change the power of the test. The efficiency decreases with the true effect size,  $E$ , because there is a greater information loss due to misclassification as  $E$  increases.<sup>2</sup> In cases where  $p^*$  is chosen to match the true population proportion (i.e.,  $a = p$  as is approximately true for the fourth and recommended threshold setting approach) then

$$\text{Classification efficiency} = \frac{(r_1 - r_0)^2}{\{[r_0(1 - r_0)p + r_1(1 - r_1)(1 - p)]E^2 + 1\}} \tag{3}$$

*The Direct Substitution Estimator*

*Properties.* The direct substitution estimator uses the  $p_i$ s as independent variables by fitting a linear regression model of the form

$$Y_i = \alpha^* + \beta p_i + \varepsilon_i^* \quad (4)$$

and tests the null hypothesis through a test of whether the regression coefficient for  $p_i$  is zero. As noted in Section A.2 of the supplementary appendix, under conditions assumed earlier, the direct substitution estimator is unbiased but has a variance inflated by misclassification. Hence, after reduction (Section A.2 of the supplementary appendix), the efficiency of direct substitution is given by

$$\text{Direct substitution efficiency} = \frac{R^2}{(1 + E^2 p(1-p)(1-R^2))} \quad (5)$$

where  $R^2 = \text{Var}(p_i)/[p(1-p)]$  is the (“pseudo”)  $R^2$  from regressing  $Z$  on  $X$ , and equals the correlation between the probabilities and  $Z_i$ . Clearly, direct substitution efficiency is  $\leq 1$  with some loss relative to full-information, except when the observed variable can be predicted perfectly. When  $R^2$  is small the loss in efficiency can be very large. If  $E$  is small, then direct substitution efficiency approximates  $R^2$ .

*Single Imputation Estimator.* A related approach sometimes used by analysts who are uncomfortable using continuous probabilities in place of truly dichotomous variables, is to create a single stochastic imputation for each  $Z_i$  by drawing a random [0–1] variable,  $W_i = 1$  with probability  $p_i$ . Tests of the hypothesis that  $\beta = 0$  are conducted using a standard  $t$ -test based on classification by  $W_i$  and using the difference in group means defined by  $W_i$  as the estimator of  $\beta$ . It can be demonstrated that this *single imputation estimator* is attenuated through random misclassification and is less efficient than direct substitution.<sup>3</sup> To achieve equal power using single stochastic imputation as is obtained by direct substitution would require a sample that is  $1/R^2$  times larger. In fact, under some circumstances, a single stochastic imputation can result in less power than classification. In general this method cannot be recommended for analysis. To be efficient, stochastic imputation must use information in the outcomes values (the  $y$ s), and multiple imputation is superior to single imputation for this purpose.

## PARTIAL INFORMATION MAXIMUM LIKELIHOOD ESTIMATION

### *Properties*

The appeal of the previous methods is that they can be implemented using standard regression package software without any custom programming of function optimization. However, computational simplicity may come at a cost of less power. Partial information maximum likelihood estimation will provide more efficient estimates of the difference in means and should provide greater power for testing hypotheses about this difference.

The PIMLE achieves greater power by incorporating information about  $Y_i$  into an implicit iterative classification of observations into either zeros or ones. To use this information about the outcomes, the PIMLE requires assumptions about the distribution of both the outcome and the missing dichotomous predictor. In this paper, we consider linear models where the following assumptions are justifiable. Let  $Y_i \sim \text{Normal}(\alpha + \beta Z_i, \sigma^2)$  and  $Z_i \sim \text{Bernoulli}(p_i)$ . Then the likelihood contribution for the  $i$ th observation is

$$f(Y_i|\alpha, \beta, p_i, \sigma^2) = p_i \phi(Y_i|\alpha + \beta, \sigma^2) + (1 - p_i) \phi(Y_i|\alpha, \sigma^2) \quad (6)$$

where  $\phi(Y_i|\alpha, \sigma^2)$  is the normal density with mean  $\alpha$  and variance  $\sigma^2$ . The likelihood is a mixture of two normal with known mixing parameters,  $p_i$ . As shown in Dempster, Laird, and Rubin (1977), obtaining the MLE for such problems requires a straightforward application of the EM algorithm. The algorithm involves a series of iteratively reweighted least squares regression problems where the weights involve the ratio of  $\phi(Y_i|\alpha + \beta, \sigma^2)$  to  $\phi(Y_i|\alpha, \sigma^2)$  with the unknown parameters set to their values from the previous iteration. The algorithm starts with a guess for the initial values and repeats iterations until the estimates converge.

The choice among classification, direct substitution, and the PIMLE may be guided by their relative power. The PIMLE will provide greater power than the other methods, but the difference in power has not been explored for this class of problems. Unlike the other estimators, the power of the PIMLE cannot be determined analytically without consideration of the distribution of  $p_i$ ; thus, in the next section we use a simulation study to calculate its relative  $T$  statistic. We compare this statistic with those of the other estimators to determine relative gains in power for the PIMLE.

*Multiple Imputation Estimator*

Another alternative that analysts might consider is what we here call the *multiple imputation estimator*—"proper" multiple stochastic imputation (Rubin, 1987). This form of stochastic imputation uses information from the outcomes when imputing the missing dichotomous predictors and provides multiple-imputed values for each  $Z_i$ . Asymptotically, as the number of multiple imputations becomes sufficiently large the method has the same power as the PIMLE.

## SIMULATION STUDY

*Design*

We conducted a simulation study comparing the efficiency of (1) the classification estimator, (2) the direct substitution estimator, and (3) the PIMLE to the full-information estimator. This simulation serves two purposes. First, it provides a sense of how the key parameters that drive the formulas for classification efficiency and direct substitution efficiency vary with characteristics of the distribution of the probabilities. Second, it serves to compare the PIMLE to the full-information estimator over realistic parameter values to estimate its efficiency as no closed-form formula exists, and then to compare this to the efficiencies of the alternative estimators.

This simulation generated data sets according to equations (1) and (4) by first simulating predicted probabilities,  $p_i$  then generating  $Z_i$ , and finally generating the outcomes,  $Y_i$ . To generate the predicted probabilities, the study used a series of six mixtures of two beta distributions parameterized by two means ( $\eta_1, \eta_2$ ), two variances ( $v_1^2, v_2^2$ ), and a mixing proportion  $\lambda$ . These six mixtures were selected to produce values of key parameters ( $r_1 - r_0$ , and  $R^2$ ) that span a realistic range of values. The values of  $\eta_1$ ,  $\eta_2$ , and  $\lambda$  used in the simulations are presented in Table A1 of the supplementary appendix. All simulations used  $v_1^2 = v_2^2 = 0.01$  because this provided interesting distributions for the probabilities without unnecessary complexity.

For each mixture, we consider effects sizes of 0.2, 0.5, and 0.8 standard deviations for generating the normal outcomes with variance one within each group. On the basis of preliminary simulations, we determined that relative efficiencies were invariant to sample sizes in the range of 200–1,600 observations. Consequently, for each design cell defined by a particular combination of effect size and distribution for  $p_i$  we generated 10,000 data sets of 400 observations each and calculated the PIMLE and full-information estimates



for each simulated data set. Using these 10,000 estimated coefficients, we calculated empirically the mean and standard error of the PIMLE and the full-information estimators, the  $T$  statistics for both estimators and the efficiency of the PIMLE as the square of the ratio of the  $T$  statistic for the PIMLE to the  $T$  for the full-information estimator. For each cell, we also determined the value of the attenuation factor ( $r_1 - r_0$ ) and  $R^2$  on the basis of the distribution for the estimated probabilities and used these parameters to calculate the classification and direct substitution efficiency using formulas (2) and (5), respectively.

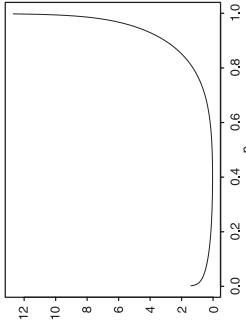
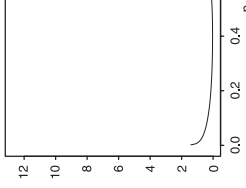
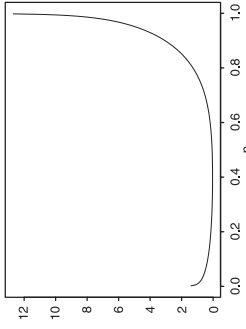
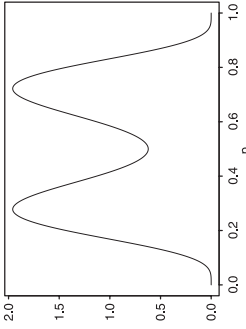
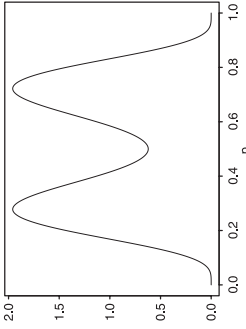
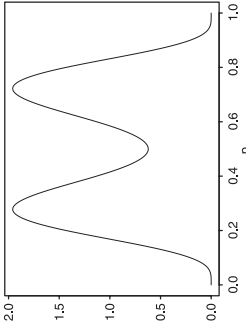
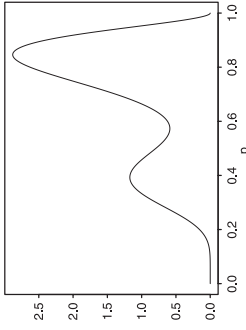
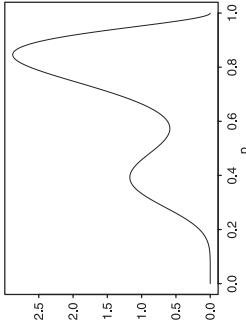
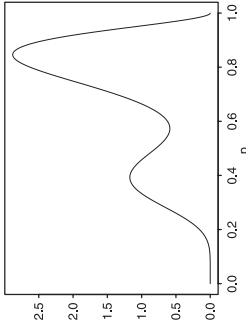
### Results

Tables 1A and 1B present the simulation results with a column for each of the three effect sizes grouped by the distribution of the predicted probabilities. For each block, we provide parameters ( $R^2$ ,  $r_1 - r_0$ , and  $p$ ) calculated analytically from the density of probabilities. There is one value of each statistic for the entire block of three effect sizes because these statistics depend only on the density of the probabilities, and not on the effect sizes or values of  $Z$  and  $Y$ . The tables also contain values of classification efficiency and direct substitution efficiency for each design cell. These values depend on both the density and the effect sizes but not the values of  $Z$  and  $Y$ . The next row provides the efficiency of the PIMLE.

The tables demonstrate that high values of  $R^2$  require extreme distributions with probabilities all very close to either 0 or 1. In Set 1, the probabilities cluster near the extremes, but  $R^2$  is only 0.46. In Set 2, the distribution of probabilities is clearly bimodal and well separated, but  $R^2$  is only 0.20. Set 3 is similar to Set 2, except for having unequal modes, something that has little effect on  $R^2$ . In the remaining cases, the probabilities are not well separated and  $R^2$  is below 0.10. The attenuation factor for classification ( $r_1 - r_0$ ) correlates highly with  $R^2$ ; scenarios with the largest  $R^2$  also have the largest values of this factor and the least attenuation.

As can be seen, the PIMLE has power equivalent to a sample 8–20 percent as large as modeling with a known predictor for five of the mixtures (and as high as 46–48 percent for the easiest of the set). Unlike the classification and substitution methods, performance of the PIMLE with respect to full information increases slightly with the effect size. As given by (5), direct substitution efficiency is essentially equal to  $R^2$ . It is also essentially equal to the efficiency of the PIMLE at effect sizes of 0.2 standard deviations for all six designs, and is 96–99 percent as efficient as the PIMLE at effect sizes of 0.5 standard deviations, with the exception of one case with an 8 percent difference

Table 1A: Calculated and Simulated Power for Classification, Direct Substitution, and PMLE Relative to Full Information\*

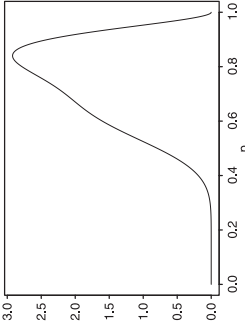
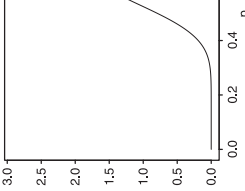
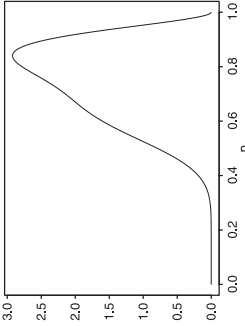
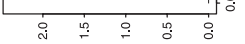
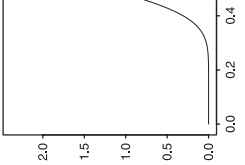
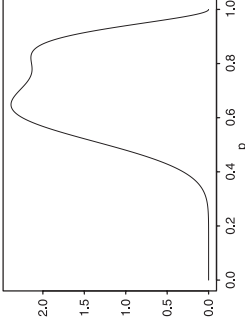
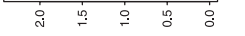
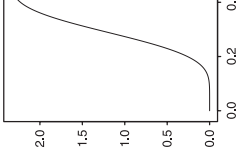
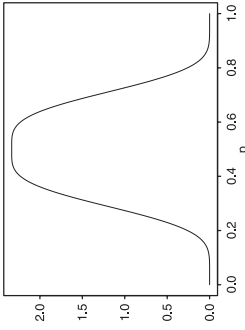
	Set 1			Set 2			Set 3		
	$R^2 = 0.46, \rho = .82$ $r_1 - r_0 = .45$			$R^2 = 0.20, \rho = .50$ $r_1 - r_0 = 0.41$			$R^2 = 0.20, \rho = .68$ $r_1 - r_0 = 0.38$		
	0.20	0.50	0.80	0.20	0.50	0.80	0.20	0.50	0.80
Effect size <sup>†</sup>									
Efficiency									
Classification	0.203	0.195	0.180	0.163	0.156	0.145	0.142	0.136	0.126
Direct substitution	0.457	0.449	0.436	0.198	0.190	0.177	0.199	0.192	0.180
PMLE (SE) <sup>‡</sup>	0.460 (0.010)	0.465 (0.007)	0.481 (0.007)	0.203 (0.006)	0.198 (0.004)	0.199 (0.004)	0.196 (0.005)	0.198 (0.004)	0.207 (0.003)
Density of predicted probabilities									

\*The probabilities that the unobserved dichotomous variables,  $Z = 1$  are a random sample from the density which is a mixture of two  $\beta$ .  $R^2$  is from the regression equation predicting the unobserved variable,  $\rho = P(Z=1)$  and  $r_1 - r_0 = P(Z=1|U=1) - P(Z=1|U=0)$ , where  $U$  is an error prone classification variable. Classification and Direct Substitution Relative Efficiency were calculated analytically, PMLE relative efficiency was calculated via Monte Carlo simulation.

<sup>†</sup>Effect size is in standard deviation units.

<sup>‡</sup>Monte Carlo standard error for simulation.

Table 1B: Calculated and Simulated Power for Classification, Direct Substitution, and PIMLE Relative to Full Information\*

	Set 4			Set 5			Set 6		
	$R^2 = 0.10, p = .74$ $r_1 - r_0 = 0.23$			$R^2 = 0.10, p = .70$ $r_1 - r_0 = 0.24$			$R^2 = 0.08, p = .50$ $r_1 - r_0 = 0.24$		
	0.20	0.50	0.80	0.20	0.50	0.80	0.20	0.50	0.80
Effect Size <sup>†</sup>									
Efficiency									
Classification	0.051	0.049	0.046	0.056	0.054	0.050	0.055	0.053	0.048
Direct substitution	0.095	0.092	0.086	0.095	0.091	0.085	0.079	0.076	0.070
PIMLE (SE) <sup>‡</sup>	0.090 (0.004)	0.096 (0.002)	0.110 (0.003)	0.094 (0.004)	0.099 (0.002)	0.111 (0.002)	0.078 (0.003)	0.076 (0.002)	0.080 (0.002)
Density of predicted probabilities									

\*The probabilities that the unobserved dichotomous variables,  $Z = 1$  are a random sample from the density which is a mixture of two  $\beta$ .  $R^2$  is from the regression equation predicting the unobserved variable,  $p = P(Z = 1)$  and  $r_1 - r_0 = P(Z = 1|U = 1) - P(Z = 1|U = 0)$ , where  $U$  is an error prone classification variable. Classification and Direct Substitution Relative Efficiency were calculated analytically, PIMLE relative efficiency was calculated via Monte Carlo simulation.

<sup>†</sup>Effect size is in standard deviation units.

<sup>‡</sup>Monte Carlo standard error for simulation.

(Set 5). By 0.8 standard deviations, direct substitution is 87–90 percent as efficient as the PIMLE, except for Set 5, where the difference is 23 percent.

As given by (2), classification efficiency is essentially equal to the square of the attenuation factor. Classification is 37–80 percent as efficient as the PIMLE, with somewhat lower relative efficiency at larger effect sizes (and 41–82 percent as efficient as direct substitution, with little variation by effect size). Compared with the other estimation methods, classification had its best relative efficiency in the symmetric case of Set 2 and its worst relative efficiency in the asymmetric case of Set 1.

## ILLUSTRATION WITH MEDICARE DATA

As an example of how this approach might be implemented, we describe an application that uses a sample of 271,479 original (Fee-for-Service) Medicare Beneficiaries surveyed as part of the 2000–2002 Consumer Assessment of Healthcare Providers and Systems (CAHPS) Medicare Fee-for-Service (MFFS) Surveys. The example considers an error-prone survey-based proxy for the true dichotomous predictor of interest, diagnosed depression. The MFFS survey includes items that allow calculation of the Mental Component Score (MCS; Ware and Kosinski 2001). As described more fully elsewhere (Health Services Advisory Group 2006), this file was merged with administrative records that provided a measure of true diagnosed depression. We used these merged files to derive an MCS-based proxy for diagnosed depression in the form of predicted probabilities. Such a measure could then be used to assess the association between depression and self-reported outcomes such as CAHPS ratings of health care, physical activity, or total utilization on any survey that includes the MCS. Notice that an analysis that uses the probability of depression is estimating the relationship between depression and an outcome of interest, which is fundamentally different than simply estimating the linear association between MCS and such an outcome.

To derive the predicted probabilities of depression, we calculated mean MCS scores by decile. Pooling adjacent deciles where the depression proportion did not decrease monotonically resulted in six ordered bins for MCS scores: < 40.5 (18.5 percent depression), 40.5–48.2 (12.6 percent depression), 48.2–52.1 (9.3 percent depression), 52.1–54.7 (6.3 percent depression), 54.7–57.8 (4.5 percent depression), and greater than 57.8 (3.2 percent depression).

The pseudo  $R^2$  for these predicted probabilities was 10 percent, so that at typical effect sizes the direct substitution approach would have the same power

as a data set with known depression that was one-tenth the size. Thus, for a given level of statistical power, sample sizes would need to be an order of magnitude larger with proxy diagnosis than with actual diagnosis. Large samples like CAHPS MFFS would provide adequate power to detect even small effects and in such cases direct substitution would likely be the method of choice. In considerably smaller samples with power to detect only moderate to large effects, PIMLE may be preferred given its somewhat greater efficiency for effect sizes in this range.

## DISCUSSION

All three methods (classification, direct substitution, and PIMLE) result in substantial loss of efficiency relative to a known predictor (the full-information estimator), due to loss of information. The  $R^2$  heuristic provides a good way to estimate information loss for study design and correction of standard error.

In most situations of interest, direct substitution will be the method of choice, as it is simpler to implement and was nearly as efficient as the PIMLE when effect sizes were 0.2–0.5 standard deviations in each of six very different distributions of predicted probabilities, with smaller losses when predicted probabilities have clear bimodality and higher values of  $R^2$ . As effect sizes reach 0.8 standard deviations, the efficiency of direct substitution fell to 87–90 percent of the efficiency of PIMLE in most cases, faring better under the same circumstances as before. It should be noted that such efficiency is likely to be adequate under these undemanding circumstances. These patterns appear to be independent of sample size and are mainly a function of the PIMLE's increasing efficiency with effect size (and to a lesser extent, a small corresponding degradation in the efficiency of direct substitution). The relative efficiency of PIMLE increases with the effect size because it implicitly uses information from the outcomes in assessing group membership. The other methods use only the information in the  $p$ s for this determination.

Classification is notably less efficient than direct substitution and PIMLE are, due to inefficient use of the  $p_i$  values. The relative performance of classification compared with direct substitution and PIMLE is insensitive to effect size and sample size in the simulated ranges.

Direct substitution is like a two-stage IV technique for endogenous or error-prone predictors. In all cases, the expected value of the independent variable is used for regression rather than the variable itself. Our finding that direct substitution results in substantial loss of efficiency compared with the

full-information estimator corresponds to the well-known result that IV can result in substantial loss of efficiency relative to experimental assignment, even when there is good correlation between the instrument and the independent variable of interest. In the IV literature, preference is now often given to maximum likelihood methods because they are more efficient. However, our results show that the PIMLE is likely to provide little additional power and larger samples and very strong instruments are the only means to obtain sufficient power for IV estimation.

Our results for classification suggest that even small measurement error in dichotomous independent variables can have an important effect on regression estimates. It is all too common that measurement error is considered only with respect to meeting a threshold of acceptability (using  $\kappa$ s, Cronbach's  $\alpha$ , or test-retest reliability), and that no further accounting for that error in sample design or analysis takes place. However, even measures that meet these thresholds can reduce power and we encourage researchers to integrate measurement error quantitatively into common metrics such as effective sample size. This way the effective sample size per unit of cost can be maximized in a way that encompasses all aspects of a study and allows trade-offs to be rationally evaluated.

We feel that the invisible nature of measurement error results in relatively too much attention being given to sampling error and sample sizes while underinvesting in the quality of the independent variables. Given that even modest reductions in the measurement error of a fairly well-estimated but error-prone measure can substantially increase the effective sample size, measurement of key predictors should be an area of intense concern, one that we hope will receive increased attention in the evaluation of proposals and research manuscripts.

This paper clearly demonstrates that measurement error can greatly reduce power for testing hypotheses and even the best methods, such as the PIMLE, recover only a fraction of the information lost through measurement error. However, the paper does have limitations. As with any simulation study, only a limited range of possible values for sample sizes, effect sizes, and the distribution of probabilities are considered. The study, however, found that results are generally invariant to sample size and that relative gains in efficiency for the MLE require rather substantial effect sizes compared with those often found in health services and similar research. Moreover, the relative efficiency of these three methods to each other is also relatively stable across the substantial range in distributions of probabilities used in this study.

The paper considers only continuous outcomes and models without additional predictor variables. If the predicted probabilities are highly collinear with other variables included in the model, this might reduce the power of models using the probabilities. Models for dichotomous outcomes require more complex techniques because substituting predicted probabilities for dichotomous independent variables in a nonlinear model such as logistic or probit regression can yield inconsistent estimates (Bhattacharya, Goldman, and McCaffrey 2006), so that more complex maximum likelihood methods might be preferable.

We also treat the predicted probabilities as known and generated from an external source, so sampling error in these values is ignored. In some cases the probabilities might be estimated from the data. For example, latent growth models (Haviland and Nagin 2005) might be used to estimate classes of growth patterns in longitudinal data and class membership might then be used to predict other outcomes or an outcome at a later time. In such a case, group membership is latent, and only the probability of group membership is available for modeling. The resultant power might be less, because there is the addition of uncertainty and misclassification due the estimation of the probabilities. Uncertainty in the  $\pi$ s can easily be incorporated into a fully Bayesian model by treating these probabilities as prior information obtained by elicitation or some another external source. Provided the prior distributions on the regression coefficients are vague, the fully Bayesian estimates of the regression coefficients should closely resemble the PIMLE except that the posterior intervals will be somewhat wider than corresponding confidence intervals because of the uncertainty in the  $\pi$ s.

The efficiency calculations assume no additional variables will be included in the models for the outcomes. Variables used to derive the probabilities are likely to be collinear with the probabilities, and this could degrade efficiency if they are included in the model. Standard formulas for the power of tests of coefficients in multiple regression models (Milton 1986) suggest that the efficiency should be reduced in proportion to the  $R^2$  from regressing the probabilities on other variables in the model.

This paper considers three common methods for using the probabilities. Other methods exist that might be more efficient where applicable. For example, if two sets of conditionally independent probabilities were available, a method described by Kane, Rouse, and Staiger (1999) would apply.

We focused on situations in which the unobserved variable is dichotomous. Predicted values for unobserved polytomous variables are also of interest. For example, race is not available in some administrative claims data,

but it might be predicted by address and surname. Extension of the classification estimator to polytomous variables would be straightforward, although classification rules would be more complicated. The extension of direct substitution would entail including in the model as independent variables the group-inclusion probabilities for all but one of the possible groups as defined by the polytomous variable. PIMLE also extends naturally. However, the relative efficiency of the methods is likely to depend on complex combinations of different types of misclassification and the multivariate distribution of the probabilities. We expect that some results for dichotomous variables will have analogs with polytomous variables (e.g., the advantages of PIMLE will be minimal when differences among group means are small relative to the variance in the outcomes) but this is an important area for future research.

## ACKNOWLEDGMENTS

The authors would like to thank Kate Sommers-Dawes and Robert Hickam for assistance with the preparation of the manuscript. Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056).

*Disclaimer:* The contents of the publication are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

## NOTES

1. In a data set with  $n$  observations, one can empirically determine the optimal  $p$  for given assumptions about the effect size because formula (2) can take on only  $n$  distinct values.
2. Tavere et al. (1995) consider the efficiency loss due to measurement error in a dichotomous independent variable for  $E=0$ . In this case, (2) simplifies to  $(r_1 - r_1)^2 a(1 - a) / \{p(1 - p)\}$ , which Tavere et al. (1995) and Schuster (2004) show equals the Cohen's  $\kappa$  statistic (1960) for the reliability of the error-prone dichotomous measure.
3. Let  $b_w$  equal the single imputation estimator, straightforward algebra shows that  $E(b_w) = R^2\beta$  and  $\text{Var}(b_w) = \sigma^2 R^2 \{1 + E^2 p(1 - p)(1 - R^2)\} / \{n\text{Var}(p_i)\}$

## REFERENCES

- Bhattacharya, J., D. Goldman, and D. McCaffrey. 2006. "Estimating Probit Models with Self-Selected Treatments." *Statistics in Medicine* 25 (3): 389-413.



- Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B: Methodological* 39 (1): 1–22.
- Fuller, W. A. 1987. *Measurement Error Models*. New York: John Wiley and Sons.
- Haviland, A. M., and D. S. Nagin. 2005. "Causal Inferences with Group Based Trajectory Models." *Psychometrika* 70 (3): 557–78.
- Health Services Advisory Group 2006. "The Evaluation of a Mental Component Summary Score Threshold for Depression Risk" Report to the Centers for Medicare and Medicaid Services, Task 5.20 Final Report [accessed November 2, 2006]. Available at [www.hosonline.org](http://www.hosonline.org)
- Kane, T., C. Rouse, and D. Staiger. 1999. "Estimating Returns to Schooling When Schooling Is Misreported." NBER #7325.
- Milton, S. 1986. "A Sample Size Formula for Multiple Regression Studies." *Public Opinion Quarterly* 50 (1): 112–8.
- Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Schuster, C. 2004. "How Measurement Error in Dichotomous Predictors Affects the Analysis of Continuous Criteria." *Psychology Science* 46 (1): 128–36.
- Steel, R. G. D., and J. H. Torrie. 1980. *Principles and Procedures of Statistics: A Biometrical Approach*. New York: McGraw Hill Book Co.
- Tavere, C. J., E. L. Sobel, and F. H. Gilles. 1995. "Misclassification of a Prognostic Dichotomous Variable: Sample Size and Parameter Estimate Adjustment." *Statistics in Medicine* 14: 1307–14.
- Ware, J., and M. Kosinski. 2001. *SF-36 Physical and Mental Health Summary Scales; A Manual for Users of Version 1*, 2nd edition. Lincoln, RI: QualityMetric Inc.

## SUPPLEMENTARY MATERIAL

The following supplementary material for this article is available online:  
Appendix A: Mathematical Derivations.

This material is available as part of the online article from <http://www.blackwell-synergy.com/doi/abs/10.1111/j.1475-6773.2007.00810.x> (this link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.