Psychology Science, Volume 46, 2004 (1), p. 128-136

How measurement error in dichotomous predictors affects the analysis of continuous criteria

CHRISTOF SCHUSTER[1]

Abstract

If the mean of a criterion variable differs across groups, measurement error in the grouping variable leads to bias in the estimated size of the effect. Similar to the classical errors-in-variables model for continuous predictors, measurement error in a dichotomous predictor leads to underestimation of the 'true' effect size. It is shown how to correct the effect size given certain characteristics of the measurement error are known.

Key words: Measurement error in predictors, predictor reliability, regression analysis

[1] Christof Schuster, Justus-Liebig-Univerisät, Fachbereich Psychologie und Sportwissenschaft, Otto-Behaghel-Str. 10, 35394, Gießen, Germany; E-mail: christof.schuster@psychol.uni-giessen.de

Introduction

When individuals are assigned to mutually exclusive categories, a certain proportion of them may be misclassified. Equivalently, one also speaks of measurement error in the classification. For instance, it is well known that classification of psychiatric patients into diagnostic categories is not perfectly reliable and therefore, the diagnoses of some patients can be expected to be incorrect. Participants in the treatment group of a clinical trial may not follow a prescribed medication plan and therefore, may be misclassified as being 'treated' when, in fact, they were not. Or, when occupational attainment of individuals is assessed by asking them to indicate the job category into which they belong, misclassification may arise because job categories sometimes do not have clear-cut boundaries.

If misclassification in a categorical variable is present, statistical analysis should account for the misclassification, otherwise the results may be biased. Of course, the amount of bias can be expected to be directly related to the extent of the misclassification. In the common situation where the dependence of a criterion variable on a predictor variable is analyzed, one has to distinguish carefully whether it is the criterion variable, the predictor variable, or both that are subject to error.

In this article, only the case of measurement error in the predictor variable will be considered; see Schuster and von Eye (2003) for a discussion of the case in which a dichotomous criterion variable contains misclassification. In particular, the focus will be on the important case of a continuous criterion variable and a dichotomous predictor, which for instance occurs when a researcher plans to compare two groups on a continuous variable, but for which individuals can not be assigned to a group with perfect reliability. In this case, the difference between the group means will be a biased estimate of the effect size and therefore, the outcomes of familiar tests on location differences, such as the two-sample *t*-test or the Wilcoxon Rank Sum test, may be questionable.

I will begin by reviewing the classical measurement error model that considers a linear model between continuous variables, both of which contain measurement error. This model is also commonly referred to as the "errors-in-variables" model. I then present a measurement error model for a continuous criterion and a dichotomous predictor. Although it is possible to maintain standard assumptions, such as uncorrelatedness of error terms, normality of error terms, and so forth, it will be seen that a dichotomous predictor variable introduces additional complexity not present in the classical measurement error model. Although this model has found some attention in the econometric and biostatistical literature (see, for example, Aigner, 1973; Christopher & Kupper, 1995; Hausman, 2001), it seems that it has been largely ignored in psychology. This is unfortunate because many psychological classifications can be expected to have less than perfect reliability. Finally, it is examined how the effect size can be corrected if the general characteristics of the measurement error are available.

The Errors-in-Variables Model for Continuous Variables

The usual linear regression model expresses the relationship between the criterion Y and the predictor X as

$$Y = \beta_0 + \beta_1 X + u, \tag{1}$$

where β_0 and β_1 are the intercept and slope parameters and u represents the residual. The residual may contain multiple sources of error, such as equation error and measurement error in the criterion. Among the standard assumptions of the linear regression model are that u has a mean of zero and that u and X are uncorrelated. Under these conditions, it is well known that the OLS estimators of the regression coefficients are unbiased. Although the linear regression model accounts for unsystematic measurement error in the criterion, the predictor is assumed to be measured without error. For social science data, the assumption of perfectly reliable predictors is often difficult to justify.

If the predictor can only be observed with measurement error, then the simple linear regression model, Equation (1), can be modified by replacing the predictor X with the measurement error-free predictor, ξ. In other words, if the predictor X can be observed with measurement error only, then one replaces the usual model for the regression of Y on X with the regression model of Y on ξ. The model then becomes

$$Y = \gamma_0 + \gamma_1 \xi + \varepsilon. \tag{2}$$

The assumptions about the residual ε are similar to the corresponding assumptions about u, the residual of the regression of Y on X. Specifically, it is assumed that ε has a mean of zero and that ε and the ξ are uncorrelated. However, because ξ can not be observed, the question arises how the γ-coefficients can be estimated. Before this issue can be addressed, additional assumptions about the measurement error in the predictor must be introduced.

It is common to assume an additive measurement error δ, that is, $X = \xi + \delta$. In addition, δ is assumed to have a mean of zero and is uncorrelated with both ξ and ε. The model of the regression of Y on ξ, Equation (2), can be expressed as

$$\begin{aligned} Y &= \gamma_0 + \gamma_1 (X - \delta) + \varepsilon \\ &= \gamma_0 + \gamma_1 X + u^*, \end{aligned} \tag{3}$$

where $u^* = \varepsilon - \gamma_1 \delta$. Because Equations (3) and (1) both express a linear model in the observable variables Y and X, these models appear to be identical. However, Equation (3) differs from the usual linear regression model because an assumption about the residual terms u of Equation (1) is not satisfied for the residual term u^* of Equation (3). Specifically, the covariance between X and u^* is equal to $-\gamma_1 \sigma_\delta^2$, and therefore, X and u^* will be correlated whenever X contains measurement error, that is, $\sigma_\delta^2 > 0$, and the mean of the criterion depends on ξ, that is, $\gamma_1 \neq 0$.

A consequence of this difference between the two models is that the γ-parameters of Equation (3) and the β-parameters of Equation (1), although conceptually closely related, are estimated by different sets of estimators. The β-parameters can be estimated easily by the usual OLS estimators (see, for instance, von Eye & Schuster, 1998). Therefore, if the relation between the β- and γ-parameters is simple, then one should be able to obtain estimators of the γ-parameters by exploiting this relationship. For the slope coefficient, γ_1, which usually is of primary interest, one can show that

$$\hat{y}_1 = \frac{1}{P_{xx}}\hat{\beta}, \tag{4}$$

where the ratio P_{xx} is commonly referred to as the predictor reliability (Cheng & Van Ness, 1999; Fuller, 1987). In this article, the predictor reliability is defined as the correlation between parallel variables, where two variables, $X_1 = \xi + \delta_1$ and $X_2 = \xi + \delta_2$ are said to be parallel, if $\mathrm{E}(\delta_1) = \mathrm{E}(\delta_2) = 0$ and $\mathrm{Var}(\delta_1) = \mathrm{Var}(\delta_2)$. Because this correlation can not exceed a value of 1.0, it follows that $\hat{\beta}_1$ can only be considered a biased and inconsistent estimator of γ_1. More specifically, $\hat{\beta}_1$ will systematically underestimate γ_1 and this bias can not be expected to decrease with increasing sample size. However, correcting the bias is in principle staightforward. If the reliability is known or if an independent and sufficiently accurate reliability estimate is available one can correct the OLS slope with the help of Equation (4).

The Errors-in-Variables Model with a Continuous Criterion and a Dichotomous Predictor

In this section, we investigate the errors-in-variables model for the situation of a continuous criterion and a dichotomous predictor containing measurement error. Therefore, in the remainder of this article X and ξ denote dichotomous variables. Although it would be desirable if one could maintain the assumptions of the errors-in-variables model for continuous predictors, this is not possible because the measurment error of the predictor neither has necessarily an expected value of zero nor is it independent of the predictor (Aigner, 1973). This is best understood by considering the joint distribution of the measurement error δ and the predictor ξ.

Table 1:
Joint distribution of 'true' predictor values and measurement error of the predictor

	$\delta = 1$	$\delta = 0$	$\delta = -1$	
$\xi = 1$	0	$(1 - \pi_1)P_\xi$	$\pi_1 P_\xi$	P_ξ
$\xi = 0$	$\pi_0(1 - P_\xi)$	$(1 - \pi_0)(1 - P_\xi)$	0	$1 - P_\xi$
	$\pi_0(1 - P_\xi)$	$1 - \pi_0(1 - P_\xi) - \pi_1 P_\xi$	$\pi_1 P_\xi$	

Let the probability of an individual belonging to the first subpopulation be denoted as P_ξ, that is, $P_\xi = P(\xi = 1)$, and let the misclassification probabilities be denoted as π_1 and π_0. Specifically, π_1 denotes the probability with which an individual belonging to the first category will be misclassified, that is, $\pi_1 = P(X = 0 \mid \xi = 1)$, and π_0 denotes the probability that an individual belonging to the other group will be misclassified, that is, $\pi_0 = P(X = 1 \mid \xi = 0)$. Using this notation, Table 1 gives the joint distribution of ξ and δ. From this joint distribution it follows that

$$E(\delta) = \pi_0(1 - P_\xi) - \pi_1 P_\xi$$
$$Cov(\xi, \delta) = -\sigma_\xi^2(\pi_1 + \pi_0),$$

where $\sigma_\xi^2 = P_\xi(1 - P_\xi)$. The second formula shows that if the true state of the individuals differs in the population and therefore, $\sigma_\xi^2 > 0$, the true predictor value and the measurement error δ will be negatively correlated.

Because the assumptions of the errors-in-variables models for continuous and dichotomous predictors differ, it is not surprising that the approaches to correcting the regression slope in the classical case do not carry over to the dichotomous predictor case. In the next section we investigate these differences.

Correcting the Regression Slope for Measurement Error when the Predictor is Subject to Misclassification

Denote the true population means of the groups studied as μ_1 and μ_0. Specifically, $\mathrm{E}(Y \mid \xi = 1) = \mu_1$ and $\mathrm{E}(Y \mid \xi = 0) = \mu_0$. If the classification is not perfectly reliable, then it is of interest to express the expectation of Y given the actual assignment group, denoted below as $E(Y \mid X = 1) = \nu_1$ and $E(Y \mid X = 0) = \nu_0$, in terms of the μ-parameters. In other words, it is of interest to determine how the μ- and ν-parameters are related.

To achieve this, one typically assumes X to not carry any additional information over and above the information that is contained in ξ. In other words, one assumes that Y and X are conditionally independent given ξ.[2] Assuming conditional independence, it is shown in the Appendix that the relationship between the μ- and ν-parameters is

$$\mu_1 - \mu_0 = \frac{1}{\tau}(\nu_1 - \nu_0), \tag{5}$$

where

$$\tau = \theta \frac{\sigma_\xi^2}{\sigma_x^2}, \tag{6}$$

[2] This assumption could be violated by a placebo effect. If $\xi = 0$ denotes the control group (absence of treatment), then the administration of an inert pill, $X = 0$, could have a positive effect that is not accounted for by the value of ξ.

$\sigma_\xi^2 = P_\xi(1-P_\xi), \sigma_X^2 = P_X(1-P_X),$ and $\theta = 1-\pi_1-\pi_0$. Note that Equation (5) is similar to Equation (4) because for dichotomous predictors, $(\mu_1-\mu_0)$ is the slope of the regression of Y on ξ and $(\nu_1-\nu_0)$ is the slope of the regression of Y on X. It can be shown that the correction factor τ can not exceed 1.0 (Aigner, 1973). This result is similar to the classical errors-in-variables model for which the correction factor, the reliability of the predictor variable, also can not exceed 1.0. In other words, $\nu_1-\nu_0$ will necessarily be smaller than $\mu_1-\mu_0$ unless there is no misclassification.

We next determine the reliability of the dichotomous predictor as the correlation between statistically equivalent (parallel) variables. The correlation between two dichotomous variables is commonly referred to as the ϕ-coefficient. If X and X' denote two parallel variables, the correlation is given by

$$\phi = \frac{P(X=1,X'=1)P(X=0,X'=0) - P(X=1,X'=0)P(X=0,X'=1)}{\sqrt{P(X=1)P(X=0)P(X'=1)P(X'=0)}} \tag{7}$$

Assuming local stochastic independence between the variables, one obtains the joint classification probabilities that are given in Table 2. With the help of this table the ϕ-coefficient can be expressed as

$$\phi = \theta^2 \frac{\sigma_\xi^2}{\sigma_X^2} \tag{8}$$

This equation has been given previously by Kraemer (1979, p. 464) as an expression for Cohen's kappa (1960).[3] Two points that are worth emphasizing follow from Equation (8).

First, the correction factor τ in Equation (6) differs from the correlation in Equation (8). More specifically, it follows from Equation (8) that $\phi = \theta\tau$. Thus, the reliability of the dichotomous predictor should not be used to correct the mean differences between groups for the misclassification because if misclassification is present, then $\phi < \tau$. As a result, using ϕ instead of τ in Equation (5) will exaggerate the true effect size $\mu_1-\mu_0$. Second, the reliability, Equation (8), is different from the variance ratio σ_ξ^2/σ_X^2, which for continuous variables is an equivalent expression for the predictor reliability. In fact, it is not difficult to verify that unlike the situation for continuous variables, this variance ratio can exceed 1.0 and therefore, is not suitable as a definition of reliability. For this reason, the reliability of the dichotomous classification has been defined as the ϕ-coefficient of *parallel* variables.

It follows from Equation (5) that the bias contained in $\nu_1-\nu_0$ can be corrected as soon as an estimate of τ is available. If one knew the reliability of the classification ϕ as well as the two misclassification probabilities π_1 and π_0, one could calculate the correction factor as $\tau = \phi/\theta$. Alternatively, one could determine estimates of the parameters π_1, π_0, and

[3] Note that if the variables are parallel, then the marginal distributions of *X* and *X'* are identical and as a result, ϕ is equivalent to Cohen's kappa (Cohen, 1960, p. 43).

Table 2:
Joint distribution of two locally independent replications X and X'

	$X' = 1$	$X' = 0$	Total
$X = 1$	$(1 - \pi_1)^2 P_\xi + \pi_0^2(1 - P_\xi)$	$(1 - \pi_1)\pi_1 P_\xi + (1 - \pi_0)\pi_0(1 - P_\xi)$	P_X
$X = 0$	$(1 - \pi_1)\pi_1 P_\xi + (1 - \pi_0)\pi_0(1 - P_\xi)$	$\pi_1^2 P_\xi + (1 - \pi_0)^2(1 - P_\xi)$	$1 - P_X$
Total	P_X	$1 - P_X$	

Table 3:
Joint distribution of locally independent replications X and X0

	$X' = 1$	$X' = 0$	Total
$X = 1$	.49	.09	.58
$X = 0$	.09	.33	.42
Total	.58	.42	

P_ξ from a latent class analysis (Clogg, 1995; Goodman, 1974; Lazarsfeld & Henry, 1968; McCutcheon, 1987), based on an independently conducted study that involves several replications of the dichotomous classification variable. These estimates could then be used together with Equation (5) to obtain an independent estimate of the correction factor τ . Such a procedure would have to ensure that τ is estimated with sufficient precision. Otherwise, the corrected mean difference could deviate considerably from the true mean difference $\mu_1 - \mu_0$. In such a setting, researchers sometimes prefer to obtain a 95% confidence interval for τ and then use its upper bound as a correction factor of the mean difference. In this way, one can be reasonably confident that one is not over-correcting the mean difference.

Example

To illustrate the effect of measurement error in a dichotomous classification variable, we assume the size of the first subpopulation as well as the two misclassification probabilities have values $P_\xi = .6$, $\pi_1 = .1$, and $\pi_0 = .1$. It then follows that $\sigma_\xi^2 = .6(.4) = .24$ and $\theta = .8$. In addition, the numerical values of the joint probabilities $P(X = i, X' = j)$ can be obtained with the help of Table 2 and are given in Table 3. Using the entries in this table together with Equation (7), one finds $\phi = .6305$. As a result, the numerical value of the correction factor is $\tau = \phi / \theta = .6305 / .8 = .79$. Alternatively, the correction factor can be calculated using Equation (6). Because $\sigma_X^2 = .58(.42) = .2436$, one can verify that $\tau = .8(.24 / .2436) = .79$.

Given the parameter values, it follows that the uncorrected mean difference, $\nu_1 - \nu_0$, equals only 79% of the true mean difference $\mu_1 - \mu_0$. In other words, the uncorrected mean difference needs to be increased by a factor of 1/.788 = 1.269. Finally note that the correction factor is very different from the ϕ -coefficient. Therefore, using the ϕ -coefficient as the correction factor would give misleading results in the sense that the true mean difference would be exaggerated.

Appendix

Using the conditional independence of Y and X, one can express the relation between ν_1 and the μ-parameters as $\nu_1 = \mu_1 P(\xi = 1 \mid X = 1) + \mu_0 P(\xi = 0 \mid X = 1)$. Similarly, one finds $\nu_0 = \mu_1 P(\xi = 1 \mid X = 0) + \mu_0 P(\xi = 0 \mid X = 0)$. Therefore, the observed effect is

$$\nu_1 - \nu_0 = \mu_1 (P(\xi = 1 \mid X = 1) - P(\xi = 1 \mid X = 0)) - \mu_0 (P(\xi = 0 \mid X = 0) - P(\xi = 0 \mid X = 1)) .$$

Because

$$\begin{gathered} P(\xi = 1 \mid X = 1) - P(\xi = 1 \mid X = 0) = (1 - P(\xi = 0 \mid X = 1)) - \\ \cdot (1 - P(\xi = 0 \mid X = 0)) = P(\xi = 0 \mid X = 0) - P(\xi = 0 \mid X = 1) , \end{gathered}$$

one finds that $\nu_1 - \nu_0 = \tau(\mu_1 - \mu_0)$, where $\tau = P(\xi = 1 \mid X = 1) - P(\xi = 1 \mid X = 0)$ is the correction factor that relates the true effect to the true observed effect. With the help of the expression $P_X = (1 - \pi_1) P_\xi + \pi_0 (1 - P_\xi)$, the correction factor τ can be re-written as

$$\begin{aligned} \tau &= P(\xi = 1 \mid X = 1) - P(\xi = 1 \mid X = 0) \\ &= P(\xi = 1, X = 1) / P_X - P(\xi = 1, X = 0) / (1 - P_X) \\ &= [(1 - \pi_1) P_\xi (1 - P_X) - \pi_1 P_X P_\xi] / (P_X (1 - P_X)) \\ &= P_\xi (1 - \pi_1 - P_X) / \sigma_\xi^2 \\ &= P_\xi (1 - \pi_1 - (1 - \pi_1) P_\xi - \pi_0 (1 - P_\xi)) / \sigma_\xi^2 \\ &= (1 - \pi_1 - \pi_2) P_\xi (1 - P_\xi) / \sigma_\xi^2 \\ &= \theta \sigma_\xi^2 / \sigma_X^2 , \end{aligned}$$

where $\theta = 1 - \pi_1 - \pi_0$ denotes the proportion of targets that are correctly classified, $\sigma_\xi^2 = P_\xi (1 - P_\xi)$ and $\sigma_X^2 = P_X (1 - P_X)$. It is useful to make the mild regularity assumption: $\pi_1 = \pi_0 < 1$ to ensure that $(1 - \pi_1 - \pi_0)$ stays positive.

References

1. Aigner, D. J. (1973). Regression with a binary independent variable subject to errors of observation. Journal of Econometrics, 1, 49–60.
2. Cheng, C., & Van Ness, J. W. (1999). Statistical regression with measurement error. London: Arnold.
3. Christopher, S. R., & Kupper, L. L. (1995). On the effects of predictor misclassification in multiple linear regression analysis. Communications in Statistics - Theory and Methods, 24 (1), 13–37.
4. Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral sciences (pp. 311–359). New York: Plenum Press.

5. Cohen, J. (1960). A coefficient of agreement for nominal tables. Educational and Psychological Measurement, 20, 37–46.
6. Fuller, W. A. (1987). Measurement error models. New York: Wiley.
7. Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61 (2), 215–231.
8. Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. Journal of Economic Perspectives, 15 (4), 57–67.
9. Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. Psychometrika, 44 (4), 461–472.
10. Lazarsfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. Boston: Houghton Mifflin. McCutcheon, A. L. (1987). Latent class analysis. Newbury Park, CA: Sage.
11. McCutcheon, A. L. (1987). Latent class analysis. Newburg Park, CA: Sage.
12. Schuster, C., & von Eye, A. (2003). Estimating the dilution effect due to unreliability of dichotomous outcome variables from kappa. Applied Developmental Science, 7 (2), 87–93.
13. von Eye, A., & Schuster, C. (1998). Regression analysis for social sciences. San Diego, CA: Academic Press.