ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Analysis of inaccurate data with mixture measurement error models

Seunghwan Park [a], Jae-Kwang Kim [b,*]

[a] Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX, 78712, USA
[b] Department of Statistics, Iowa State University, Ames, IA, 50011, USA

ARTICLE INFO

Article history:
Received 24 April 2017
Accepted 24 July 2017
Available online xxxx

AMS 2000 subject classifications:
62D05
62-07

Keywords:
Fractional imputation
Missing data
Survey sampling

ABSTRACT

Measurement error, the difference between a measured (observed) value of quantity and its true value, is perceived as a possible source of estimation bias in many surveys. To correct for such bias, a validation sample can be used in addition to the original sample for adjustment of measurement error. Depending on the type of validation sample, we can either use the internal calibration approach or the external calibration approach. Motivated by Korean Longitudinal Study of Aging (KLoSA), we propose a novel application of fractional imputation to correct for measurement error in the analysis of survey data. The proposed method is to create imputed values of the unobserved true variables, which are mis-measured in the main study, by using validation subsample. Furthermore, the proposed method can be directly applicable when the measurement error model is a mixture distribution. Variance estimation using Taylor linearization is developed. Results from a limited simulation study are also presented.

1. Introduction

Measurement error indicates the difference between measured (or observed) value and the true value of a variable, a problem often encountered in survey data. A well-known example of measurement error *attenuation* problem occurs when attempting regression analysis without accounting for the error, in which case the regression coefficient tends to shrink towards zero. Attenuation inevitably weakens the statistical power of the analysis and may result in unwarranted conclusions; therefore, developing statistical methodology to help overcome the problem is an important research area in statistics. Some existing research relevant to measurement error includes Fuller (2009) and Carroll, Ruppert, Stefanski, and Crainiceanu (2006).

One way to account for measurement error is to utilize additional, separate data, referred to as validation sample, to obtain information of the true value of the variable. Obtaining the distribution of measurement error from a sample distinct from the original sample is called external calibration; randomly selecting partial data from the original sample and obtaining the distribution of measurement error from this validation sample is called internal calibration (Guo & Little, 2011). For example, let x be the auxiliary variable and y be the dependent variable. Suppose that we observe $z = x + u$ instead of x in the original sample. Here, u is a random variable representing measurement error. In this case, the data structure of internal and external calibration study can be understood as Table 1.

In the case of internal calibration, the original sample can be decomposed into validation sample and non-validation sample, depending on whether the true value x is observed or not. As such, internal calibration can be understood as a

* Corresponding author.
E-mail address: jkim@iastate.edu (J.-K. Kim).

http://dx.doi.org/10.1016/j.jkss.2017.07.002
1226-3192/

Table 1
Internal calibration study and external calibration study.

Internal calibration study	y_i	x_i	z_i
Sample A	O	X	O
Sample B	O	O	O
External calibration study	y_i	x_i	z_i
Sample A	O	X	O
Sample B	X	O	O

Sample A is an original sample and Sample B is a validation sample. O: observed, X: missing.

typical study of two-phase sampling. In fact, it can be understood as observing (y, z) from the original sample in the first-phase sample and observing (y, x, z) in the second-phase sample.

Two-phase sampling is a cost-efficient method first implemented by Neyman (1938) who drew samples across two phases: conducting a low-cost measurement in the first phase and extracting a partial sample and conducting a more delicate measurement in the second phase. This two-phase sampling is also known as double sampling and is frequently used in epidemiology, econometrics, biostatistics, and forestry. Case-control study often used in epidemiology is also a type of two-phase sampling. For example, Scott and Wild (1986) used a two-phase sampling design in their study on driving under alcohol influence, where the first-phase sample consisted of measurements taken from breathe exhalation and the second-phase sample contained measurements from more detailed blood tests. Theoretical studies related to two-phase sampling can be found in Lawless, Kalbfleisch, and Wild (1999), Breslow, McNeney, and Wellner (2003), and Chen and Rao (2007). Moreover, research that provides data analysis on measurement errors through two-phase sampling includes Carroll et al. (2006), Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995), Pepe and Fleming (1991), Carroll and Wand (1991), and Reilly and Pepe (1995).

External calibration requires additional assumptions regarding model identifiability. In this case, a frequently used assumption is non differential measurement error assumption, which is written as follows:

$$f(y|x, z) = f(y|x). \tag{1}$$

In other words, when true value x is known, additional information on z becomes unnecessary in predicting y. Variable z is also called the surrogate variable of x in some literature. External calibration also depends on the availability of an external sample distinct from the original sample, accurate measurement of true value x in the external sample, and the assumption that the population distribution of the external sample is equivalent to the distribution of the original sample. This assumption is called transportability. It is only under the premise of transportability that using an external sample to calibrate the measurement error of the original sample is statistically meaningful (see Carroll and Stefanski (1994)).

This study reviews the statistical methodology that can be applied to internal and external calibration, makes suggestions of more realistic and statistically efficient methodology, and applies it to Korean Longitudinal Study of Aging (KLoSA). Xu, Kim, and Li (2017) develop a semiparametric approach in this setup. The main contribution of this paper is to propose an efficient imputation approach to adjust for measurement bias using a finite mixture model for measurement error model. The proposed method is to create imputed values of the unobserved true variables, which are mis-measured in the original sample, by using a relatively small validation subsample. Further, we consider a new class of estimators for the regression coefficients based on expected estimating equations using the conditional distribution of latent variables given observed data. The suggested method, based on maximum likelihood estimation, not only makes way for more efficient estimation but also applies well to various types of measurement error models. The method used in this study is the parametric fractional imputation method proposed by Kim (2011) and Kim and Hong (2012). Variance estimation using Taylor linearization is developed.

This paper is organized as follows. The description of KLoSA is presented in Section 2. In Section 3, basic setup is described in the context of a measurement error model. Proposed method is presented in Section 4, where we propose parameter estimation procedures for internal and external calibration studies. In Section 5, results from two limited simulation studies are presented. Application of the proposed method to KLoSA is described in Section 6. Concluding remarks are made in Section 7.

2. Data description

This paper is motivated by a real data example from Korean Longitudinal Study of Aging (KLoSA) survey, organized by Korea Labor Institute (KLI). KLoSA is a longitudinal survey conducted every two years from 2006 surveying South Korean adults aged 45 or over. First panel (2006) contains data on the heights and weights of 10,254 individuals obtained through survey questionnaires. For a subsample of 527 individuals from the first panel, actual biometric data of height and weight were obtained. In other words, we have data on the heights and weights of 527 individuals, both in terms of their responses and actual measurements. Variable of interest, denoted as y, is whether or not an individual has high blood pressure. Many

Please cite this article in press as: Park, S., Kim, J.-K., Analysis of inaccurate data with mixture measurement error models. Journal of the Korean Statistical Society (2017), http://dx.doi.org/10.1016/j.jkss.2017.07.002.