International Journal of Statistical Distributions and Applications
2023; 9(1): 35-40
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20230901.14
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Two Sample Approaches to Regression Calibration for Measurement Error Correction

Samuel Joel Kamun[1], Cornelious Nyakundi[1], Richard Simwa[2]

[1]Department of Mathematics and Actuarial Sciences, Catholic University of Eastern Africa, Nairobi, Kenya
[2]Department of Account, Finance and Economics, School of Business, KCA University, Nairobi, Kenya

Email address:

samuelkamun@gmail.com (Samuel Joel Kamun), nyakundicornelious@yahoo.com (Cornelious Nyakundi), rsimwa@kca.ac.ke (Richard Simwa)

To cite this article:

Samuel Joel Kamun, Cornelious Nyakundi, Richard Simwa. Two Sample Approaches to Regression Calibration for Measurement Error Correction. *International Journal of Statistical Distributions and Applications*. Vol. 9, No. 1, 2023, pp. 35-40.
doi: 10.11648/j.ijsd.20230901.14

Received: February 1, 2023; **Accepted**: March 1, 2023; **Published**: March 9, 2023

Abstract: The goal of this work is to create methods for enhancing measurement error using regression calibration as a strategy by combining two samples, thereby increasing the relative efficiency of linear regression models. Because two or more samples are more likely to provide an accurate representation of the population than a single sample under inquiry, utilizing two samples in regression calibration is likely to produce a realistic depiction of what the actual population is when error-free. This study has generated independent estimates from two samples and combined them with weights equal to the inverse of their estimated probabilities of sample inclusion. It has also integrated two data sets into a single data set and suitably adjusted the weights on each sampled unit. The regression calibration method is most commonly used to correct predictor-response bias caused by variable measurement imperfections. Because of its simplicity, this method is often used. The fundamental principle behind regression calibration is to estimate the conditional expectation of a genuine response, given predictors measured with error and other covariates supposed to be measured without error. The predicted values are then estimated and used to assess the relationship between the response and an outcome in place of the unknown genuine response. Further information on the unobservable true predictors is required by the regression calibration program. This data is frequently obtained from a validation study that employs unbiased measurements for genuine predictors. This study has employed and compared the results obtained from the two sample approaches. Measuring errors can be produced by a variety of sources, including instrument error, laboratory error, human error, problems in documenting or executing measurements, self-reporting errors, and natural oscillations in the underlying amount. Covariate measurement error has three effects: In addition to hiding the properties of the data, which makes graphical model analysis difficult, it produces bias in parameter estimates for statistical models, resulting in a sometimes significant loss of power for detecting fascinating correlations between variables. The two sample approaches employed by the study have yielded acceptable results.

Keywords: Multiple Samples, Regression Calibration, Population, Error Free, Inclusion Probabilities

1. Introduction

1.1. Background

In this work, two-sample technique was employed to improve the efficiency of the regression model's coefficients through modeling and measurement error correction. In some ways, measurement error is the source of all statistical issues. When one or more variables in an interest model cannot be precisely measured, measurement error occurs. Such errors can occur for a variety of reasons, the most common of which being instrument and sample error.

1.2. Measurement Error in Exposure Variables

Measurement inaccuracy in exposure factors is well documented in a variety of research disciplines. Measurement error is defined as the difference between a variable's true and measured values [15]. Memory bias can occur when investigations are conducted in the past and

require a researcher to recall and record previous experiences. Measuring errors in research can also come from biological variations and equipment faults in laboratory testing. Assessing exposure accuracy has always been a challenge in research relating exposures to health outcomes [13].

This study focuses on the bias in exposure-outcome correlations when exposure variables are measured with errors. Because of the many exposures and associated inaccuracies, the exposure-outcome relationship may be biased in any way [5]. The presence of measurement error in the exposure problem has sparked a wave of technique research, initially focusing on understanding the effects of measurement error on the relationship between exposure and outcome and, more recently, on developing statistical approaches to correct for exposure measurement error [1, 4, 6].

2. Objectives

2.1. General Objective

Two Sample Approaches to Regression Calibration for Measurement Error Correction.

2.2. Specific Objective

1) Regression calibration is used to adjust for measurement inaccuracy in linear models.
2) Create two sample models for better measurement error adjustments.

3. Common Methods to Correct for Measurement Error

This section lists the five most commonly used methodologies for bias adjustment and focuses on only one: regression calibration, the likelihood method, simulation extrapolation (SIMEX), Bayesian methods, and multiple imputation. This paper has concentrated on just one, describing regression calibration as a method for measurement error correction and how to improve its effectiveness by using numerous samples.

3.1. Regression Calibration

The regression technique with calibration is most commonly used to adjust for bias in the predictor-response relationship caused by measurement imperfections in the variables [2, 8, 9, 11, 14]. This method is popular since it is simple. The fundamental principle behind regression calibration is to estimate the conditional expectation of a genuine response, given predictors measured with error and other covariates supposed to be measured without error. The predicted values are then estimated and used to assess the relationship between the response and an outcome in place of the unknown genuine response.

Further information on the unobservable true predictors is required by the regression calibration program. This data is frequently obtained from a validation study that employs unbiased measurements for genuine predictors. A validation study is usually smaller than the initial study and may include a random sample of the subjects from the first study [9]. When employing regression calibration, the measurement error in the predictors is often assumed to be non-differential [10]. In most cases, the technique produces consistent estimators of the association parameter [2].

3.2. Measurement Error and Its Effects

3.2.1. Measurement Error in Exposures

Blood pressure, biomarker readings, weight or height, calorie consumption, and levels of physical activity are examples of wrongly measured exposures.

3.2.2. Sources of Measurement Error

Instrument error, laboratory error, human error, faults in documenting or executing measurements, self-reporting errors, and natural oscillations in the underlying amount can all cause measurement errors.

3.2.3. Set-up Notation

Y: Outcome
X: True exposure
X*: Measured exposure
U: Error

$$\begin{array}{c} U \\ \downarrow \\ X^* \\ \uparrow \\ X \to Y \end{array} \tag{1}$$

The Linear regression model is given by

$$Y = \beta_0 + \beta_X X + \varepsilon \tag{2}$$

The following three impacts are caused by covariate measurement errors: In addition to hiding data properties, which makes graphical model analysis difficult, it produces bias in parameter estimates for statistical models, resulting in a sometimes significant loss of power for discovering fascinating correlations between variables.

3.3. Classical Measurement Error

Definition: The classical measurement error model is defined as

$$X^* = X + U \tag{3}$$

where U has a mean of 0 and a variance of σ^2_U. According to the paradigm, X is an impartial measure of X. We could get close to the truth, X, if we got multiple measurements of X from the same person and averaged them.

3.3.1. The Effects of Classical Measurement Error

Equation represents the linear regression model with X. (2),

$$Y = \beta_0 + \beta_X X + \varepsilon$$

Using the linear regression model with X*

$$Y = \beta_0^* + \beta_X^* X^* + \varepsilon \quad (4)$$

The impacts of classical measurement error include a distorted (attenuated) estimate of the slope of the connection and a loss of power to detect a link between variables.

3.3.2. Measuring the Bias Caused by Measurement Error

The linear regression model that employs X, equation (2),

$$Y = \beta_0 + \beta_X X + \varepsilon$$

The linear regression model using X*, equation (3),

$$Y = \beta_0^* + \beta_X^* X^* + \varepsilon$$

The factor of attenuation

$$\beta_X^* = \left\{ \frac{\text{var}(X)}{\text{var}(X) + \text{var}(U)} \right\} \beta_X = \lambda \beta_X \quad (5)$$

The attenuation factor λ must be determined to account for the influence of the classical measurement error. An external study, a validation study in which X is seen alongside X* in a sample of study participants, and a replication study in which repeated assessments of X* are collected in a subset of research participants are all required for the study to estimate λ.

3.3.3. Including Adjustment Variables

Z is a perfectly measured covariate in a linear regression model with adjustment factors,

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon \quad (6)$$

$$Y = \beta_0^* + \beta_X^* X^* + \beta_Z^* Z + \varepsilon \quad (7)$$

Unlike errors in X, which cause *X to be attenuated (biased towards the null), errors in Z cause *Z to be skewed in any direction. The study emphasizes the significance of inaccuracies in confounding variables.

3.3.4. Error in a Number of Explanatory Variables

The linear regression model of interest is represented by equation (6),

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon$$

where X is measured with classical error and we observe X* rather than X. Z is also measured with classical error, and we observed Z* instead of Z. Using an error-prone metric and a linear regression model, equation (7),

$$Y = \beta_0^* + \beta_X^* X^* + \beta_Z^* Z^* + \varepsilon$$

Where β*X and β*Z may be biased in any direction.

3.3.5. The Standard Regression Calibration Setting

The study is interested in a regression with an outcome Y that has at least one error-prone X and possibly other precise covariates, a reasonable prediction model for unobserved X based on observed covariates, and data that informs the structure of the measurement error, such as modeling E[X|X*, Z], where the error in X* is independent of Y and also independent of (X, Z).

3.3.6. Performing Regression Calibration (RC)

The investigation begins by fitting the model for unobserved X to observed data: $\hat{X} = E[X|X^*, Z]$. Second, the study replaces the unobserved X in the outcome regression of interest with $\hat{X}$, and third, the study corrects standard errors in the outcome model caused by having to estimate $\hat{X}$.

The bootstrap and sandwich SE methods are two ways of obtaining SE for parameter estimates in a final result regression model.

3.3.7. Regression Calibration (RC) for Linear Regression

Assume that, in equation (6),

$$Y = \beta_0 + \beta x X + \beta z Z + \varepsilon$$

and

$$X^* = \alpha_0 + \alpha_x X + \alpha_z Z + U \quad (8)$$

then

$$E[Y|X^*, Z] = E_{X|X^*, Z}[E(Y|X^*, Z)|X] = E_{X|X^*,Z}[E(Y|Z, X)]$$

$$= E_{X|X^*,Z}[\beta 0 + \beta x X + \beta z Z]$$

$$= \beta_0 + \beta x E[X|X^*, Z] + \beta z Z \quad (9)$$

Finally, regress Y on E[X|X*, Z] and Z to obtain the correct β coefficients. Where E[X|X*, Z] is the calibrated exposure.

4. Two Sample Approach for Improving the Efficiency of Measurement Error Correction

This work focuses on a two-sample strategy for improving the effectiveness of measurement error regression calibration. Assume two distinct samples and acquire relevant data about a single population, U. The paper presented three approaches for integrating data from the two samples in order to generate a single set of estimates of a population quantity or population parameter.

A general solution to this problem is to obtain independent estimates from two samples and combine them with weights that are the inverse of their estimated variances [12] and their references. Another method is to combine two data sets into a single data set and alter the weights on each sampled unit appropriately [7].

As a result, there are now two identically sized samples of distinct or identical types that combine to form a bigger sample $s = s_1 \cup s_2$. Because of overlap, the number of different units in the combined sample would likely be less than the total.

4.1. Design-Based Approaches

4.1.1. Blended Methodology I

Get separate estimates for each sample and combine them by the inverse of their estimated variances.

$$\hat{T}_{mix,I} = \frac{\hat{T}_1 * E(y_1) + \hat{T}_2 * E(y_2)}{\left(E(y_1) + E(y_2)\right)} \tag{10}$$

Individual estimators $\hat{T}_k$ suited for the respective samples, such as their corresponding Horvitz-Thompson estimators, would be used in this case. The blended estimator was then calculated based on the predictor variables of the regression function, as illustrated below:

$$\hat{T}_{mix} = E\left(\frac{\hat{T}_1 * E(y_1) + \hat{T}_2 * E(y_2)}{\left(E(y_1) + E(y_2)\right)} \middle| x_1, x_2, x_3, x_4 \right) \tag{11}$$

4.1.2. Blended Methodology II

Obtain independent estimates for each sample and weight them together by the inverse of their estimated variances, with the estimate with the lowest variance receiving the most weight.

$$\hat{T}_{mix,II} = \frac{\left(\hat{v}_{l\arg e}\hat{T}_{small} + \hat{v}_{small}\hat{T}_{l\arg e}\right)}{\left(\hat{v}_{l\arg e} + \hat{v}_{small}\right)} \tag{12}$$

Individual estimators $\hat{T}_k$ suited for the respective samples, such as their corresponding Horvitz-Thompson estimators, would be used in this case. The blended estimator was then calculated based on the predictor variables of the regression function, as illustrated below:

$$\hat{T}_{mix,II} = E\left(\left(\frac{\hat{v}_l\hat{T}_s + \hat{v}_s\hat{T}_l}{\hat{v}_l + \hat{v}_s} \right) \middle| x_1, x_2, x_3, x_5 \right) \tag{13}$$

4.2. Using Overall Inclusion Probabilities

To employ design-based techniques and make an estimate on the combined data s, we must first get the overall inclusion probabilities $\pi_i^* = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$. They can be used with a Horvitz-Thompson estimator, a Hajak estimator, or a model-assisted estimator. For example, the Horvitz-Thompson estimator would be $\hat{T}_{\pi^*} = \sum_{i \in s_1 \cup s_2} w_i^* y_i$ with $w_i^* = \pi_i^{*-1}$ should be noticed that units that occur in both samples ("overlaps units") appear in the expression only once. This presupposes that duplicates can be identified.

Weights, depending on their design, generally serve a dual purpose. They are expressly chosen because they produce desirable sampling features, such as the average of all possible samples used in the design being the target ("unbiasedness") or being sufficiently close to it ("near-unbiasedness"), and so on. They appear appropriate for the given sample insofar as they provide a fair mechanism for the units in the sample to appropriately reflect the population. As a result, sampling weights have two unique characteristics: (a) their sampling properties across prospective samples, and (b) their representativeness for this sample within this population.

4.3. Model-Assisted Semi-Parametric Regression

This study has modified the model-assisted regression of [3] for non-parametric regression estimation using model-assisted semi-parametric regression estimation. According to [3], the model-assisted estimator would be a design-unbiased estimator of

$$\hat{T}^* = \sum_s \frac{y_i - m\left(x_j\right)}{\pi_i} + \sum_U m\left(x_j\right)$$

if the actual means m(x_j) were established for $j \in U$. For this study, the model-assisted estimator would be a designed-unbiased estimator of (Option A)

$$\hat{T}^* = \sum_s \frac{y_i - \left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right)}{\pi_i} + \sum_U \left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right) \tag{14}$$

The genuine objectives, the y_i, and the mean values of $\left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right)$, which are only predicted to be close to the yi, differ from each other. The weighted-up residual adjustment in the first term takes this into account.

The primary concept is to estimate the $\left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right)$ using regression and then plug them into the preceding formula, equation (14),

$$\hat{T}^* = \sum_s \frac{y_i - \left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right)}{\pi_i} + \sum_U \left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right).$$

Option B is a variant of option A, as given in equations (2) & (3) below:

$$\hat{T} = sqrt\left(\sum_{i \varepsilon s}\left(y_i - \beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right) * \sum_{i \varepsilon s}\left(y_i - \beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right)\right) + \sum_{j \in U \setminus s}\left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right) \tag{15}$$

and

$$\hat{T}_{\pi^*} = sqrt\left(\sum_{i \in s}\left(\frac{y_i - \left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right)}{\pi_i}\right) * \sum_{i \in s}\left(\frac{y_i - \left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right)}{\pi_i}\right)\right) + \sum_{j \in U}\left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon\right) \tag{16}$$

4.4. The Horvitz-Estimator

Below is the Horvitz-Thompson estimator which for our study serves as a reference model with which the performance of models (1), (2) and (3) have been compared based on coefficient of determination, sample bias and standard error.

$$\hat{T}_{\pi^*} = \sum_{i \in s_1 \cup s_2} w_i^* y_i \tag{17}$$

The study equally modified the Horvitz-Thompson model to obtain models which outperform it. Below is a modified model of the Horvitz-Thompson model:

$$\hat{T}_{\pi^*} = E\left(\sum_{i \in s_1 \cup s_2} w_i^* y_i \middle| x_1, x_2, x_3, x_4 \right) \tag{18}$$

4.5. Simulation Study

The simulation analysis revealed that the model-assisted semi-parametric estimators outperformed the rival non-model-assisted Horvitz-Thompson estimators by a wide margin.

4.6. Two Sample Estimators and Their Equations

Table 1 summarizes the two sample estimators built as well as the two sample Horvitz-Thompson estimator.

Table 1. *Summary of two sample estimators constructed and their equations.*

Estimator	Formula	Comment
Blended Methodology I, BMI	$\hat{T}_{mix} = E\left(\frac{\hat{T}_1 * E(y_1) + \hat{T}_2 * E(y_2)}{(E(y_1) + E(y_2))} \middle| x_1, x_2, x_3, x_4 \right)$	
Blended Methodology II, BMII	$\hat{T}_{mix,II} = E\left(\left(\frac{\hat{v}_l \hat{T}_s + \hat{v}_s \hat{T}_l}{\hat{v}_l + \hat{v}_s} \right) \middle| x_1, x_2, x_3, x_5 \right)$	$\hat{v}_l$= Larger variance, v_s= Smaller variance
Semi-parametric regression I, SPRI	$\hat{T} = sqrt\left(\sum_{i \in s} \left(y_i - \beta_0 + \beta_{x_j} x_j + \varepsilon \right) * \sum_{i \in s} \left(y_i - \beta_0 + \beta_{x_j} x_j + \varepsilon \right) \right) + \sum_{j \in U \setminus s} \left(\beta_0 + \beta_{x_j} x_j + \varepsilon \right)$ $\hat{T}_{SPRI} = E\left(\hat{T} \middle| x_1, x_2, x_3, x_4 \right)$	
Semi-parametric regression II, SPRII	$\hat{T}_{\pi^*} = sqrt\left(\sum_{i \in s} \left(\frac{y_i - \left(\beta_0 + \beta_{x_j} x_j + \varepsilon \right)}{\pi_i} \right) * \sum_{i \in s} \left(\frac{y_i - \left(\beta_0 + \beta_{x_j} x_j + \varepsilon \right)}{\pi_i} \right) \right) + \sum_{j \in U} \left(\beta_0 + \beta_{x_j} x_j + \varepsilon \right)$ $\hat{T}_{SPRII,\pi^*} = E\left(\hat{T}_{\pi^*} \middle| x_1, x_2, x_3, x_4 \right)$	$\pi_i = \pi^*$ $\pi_i^* = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$
Semi-parametric Conditional Regression, SPCR	$\hat{T}_{SPCR,\pi^*} = E\left(\sum_{i \in s_1 \cup s_2} w_i^* y_i \middle| x_1, x_2, x_3, x_4 \right)$	$\pi_i = \pi^*$ $\pi_i^* = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$
Horvitz-Thompson, HT	$\hat{T}_{HT,\pi^*} = \sum_{i \in s_1 \cup s_2} w_i^* y_i$	$w_i^* = \pi_i^{*-1}$ $\pi_i^* = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$

5. Conclusion

By combining estimates from two samples and estimating the coefficients of the regression function, this study has developed four approaches that increase the effectiveness of the regression calibration method. The results of these approaches are significantly better than the estimates from the weighted Horvitz-Thompson model, which served as the study's reference model.

References

[1] Agogo, G. O., van der Voet, H., van't Veer, P., Ferrari, P., et al. (2014). Use of Two-Part Regression Calibration Model to Correct for Measurement Error in Episodically Consumed Foods in a Single-Replicate Study Design: EPIC Case Study. PLoS ONE 9 (11): e113160. doi: 10.1371/journal.pone.0113160.

[2] Brazzale, A. R. and Guolo, A. (2008). A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. Stat.med, vol. 27, issue 19, pp. 3755-3775.

[3] Breidt, F. J. and Opsomer, J. D. (2000), Local polynomial regression estimators in survey sampling, *Annals of Statistics*, 28, 1026-1053.

[4] Buonaccorsi, J. P. (2010). Measurement Error: Models, Methods and Application. Chapman Hall/CRC.

[5] Buzas, J. S., Stefanski, L. A. and Tosteson, D. (2014). Measurement Error. In: Ahrens, W., Pigeot, I (eds). Handbook of Epidemiology. Springer, New York, NY. https://doi.org/10.1007/978-0-387-09834-0_19.

[6] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (2006). Measurement Error in Nonlinear Models. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781420010138.

[7] Dorfman, A. H. (2008), The two sample problem, *Proceedings of the Joint Statistical Meetings, Section of Survey Research Methods*. Journal of the American Statistical Association, 87, 998-1004.

[8] Fraser, G. E. and Stram, D. O. (2001). Regression Calibration in studies with correlated variables measured with error. Americal Journal of Epidemiology, vol. 154, issue 9, pp. 836-844.

[9] Freedman, L. S., Midhune, D., Carroll, R. J. and Kipnis, V. (2008). A Comparison of regression Calibration, Moment Reconstruction and imputation for adjusting for covariate measurement error in regression. Stat. Med. 27 (25): 5195-5216; doi: 10.1002/sim3361.

[10] Keogh, R. H. and White, I. R. (2014). A toolkit for Measurement Error correction, with focus on nutritional epidemiology. Stat.Med. 33 (12): 2135-55.

[11] Masser, K. and Natarajan, L. (2008). Maximum Likelihood, Multiple imputation and regression calibration for measurement error adjustment. Stat.Med. vol. 27, issue 30, Annual Conference of the International Society for Clinical Biostatistics, pp 6332-6350.

[12] Merkouris, T. (2004), Combining independent regression estimators from multiple surveys, *Journal of the American Statistical Association*, 99, 1131-1139.

[13] Rothman, K. J., Greenland, S. and Lash, T. L. (2008). Modern Epidemiology. Wolters Kluwer|Lippincott Williams & Williams.

[14] Spiegelman, D. (2013). Regression Calibration in air pollution Epidemiology with exposure estimated by spatio-temporal modelling. Environmetrics, 24 (8), 521. https://doi.org/10..1002/env.2249.

[15] THOMAS, d., Stram, D. and Dwyer, J. (1993). Exposure Measurement Error: Influence on Exposure-Disease relationships and Methods of correction. Annu. Rev. Publ. Health. 14; 69-93.