International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2023; 8(2): 46-50

https://www.mathsjournal.com
Received: 18-01-2023
Accepted: 23-02-2023

Samuel Joel Kamun
Department of Mathematics and Actuarial sciences, Catholic University of Eastern Africa, Nairobi, Kenya

Cornelious Nyakundi
Department of Mathematics and Acturial Science, Catholic University of Eastern Africa, Nairobi, Kenya

Richard Simwa
School of Business, Department of Accounting, Finance and Economics, KCA University, Kenya

Corresponding Author:
Samuel Joel Kamun
Department of Mathematics and Actuarial sciences, Catholic University of Eastern Africa, Nairobi, Kenya

A comparison of two sample approaches to regression calibration for measurement error correction

Samuel Joel Kamun, Cornelious Nyakundi and Richard Simwa

DOI: https://doi.org/10.22271/maths.2023.v8.i2a.945

Abstract
This study compares ways for improving regression calibration. This is a method for combining two samples in order to reduce measurement error and improve the relative efficiency of linear regression models. Since two or more samples are more likely than a single sample to accurately represent the population under study, two samples are used in regression calibration to produce a realistic picture of the actual population. In this investigation, we compared independent estimates derived from two samples using a weight equal to the reciprocal of the estimated sampling probability. The study also examined the estimations produced after combining the two datasets into one, and modified the weight of each sample unit accordingly. The most typical application of regression calibration methods is to account for bias in projected responses induced by measurement inaccuracies in variables. Because of its simplicity, this method is commonly utilized. The conditional expectation of the genuine response is estimated using regression calibration, given that the predictor variables are measured with error and the other covariates are assessed without error. Instead of the unknown genuine response, predictors are estimated and used to examine the link between response and result. Regression calibration programs necessitate extensive knowledge of unobservable true predictors. This information is frequently collected from validation studies that employ unbiased measurements of true predictors. The results of two sample strategies were employed and compared in this study. Device fault, laboratory mistake, human error, difficulty documenting or completing measurements, self-reported errors, and intrinsic vibrations of the underlying instrument can all cause measurement inaccuracies. Covariate measurement error has three consequences: In addition to obscuring data features and making graphical model analysis more difficult, estimates of statistical model parameters might be skewed, and effectiveness in detecting correlations between variables can be severely impaired. This study's two sampling procedures produced satisfactory results.

Keywords: Two samples, regression calibration, population, error free, inclusion probabilities

1. Introduction
The enhanced relative efficacy of statistical analyses obtained by modeling and resolving measurement error using regression calibration by merging estimates from two samples is examined in this work. All statistical errors can be traced back to imperfections in measurement. This is referred to as "measurement error," and it occurs when one or more variables in an interest model cannot be measured consistently. Such errors can occur for a variety of reasons, the most common of which are sample and instrument faults.

1.1 Exposure variable measurement error
In a variety of study disciplines, measurement inaccuracy in exposure factors has been frequently shown. Measurement error is defined as the difference between a variable's true and measured values [8]. Memory bias can occur while doing historical studies that necessitate the researcher recalling and documenting earlier experiences. Biological variations and laboratory equipment faults can also create measurement errors in a study. Assessing exposure accuracy has long been an issue in research on exposures and health effects [7].
This study investigates the bias in exposure-outcome correlations that occurs when exposure variables are recorded incorrectly. Due to the various exposures and accompanying

inaccuracies, the exposure-outcome connection may be skewed in any way [5]. The presence of measurement error in the exposure problem has sparked a wave of technique research, with the initial focus 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, [3, 8, 4].

1.2 The two sample approach for improving the efficiency of measurement error correction

In this study, two-sample approaches for enhancing measurement error regression and calibration accuracy were examined. Assume two distinct samples and gather pertinent information about the population as a whole. The study offered four approaches for integrating data from the two samples in order to provide a single set of more accurate estimates of a population number or population characteristic.

In general, this issue has been solved by combining independent estimates from the two samples and weighting them by the reciprocal of their calculated variances [6]. The two data sets might alternatively be combined and the weights on each sampled unit adjusted proportionately [5].

The study also generated a variation of the Horvitz-Thompson estimate by taking the square root of the conditional expectation of the product of the estimates from the two samples based on the predictor variables of a regression function.

2. Objectives

2.1 General Objective

A comparison of two sample approaches to regression calibration for measurement error correction.

2.2 Specific Objective

Comparison of four approaches using coefficients of determination, small sample bias and standard error.

3. Design-based approaches

3.1 Blended Methodology I

The blended model for the study is a weighted regression of the conditional expectations of predictor variables, where the weights are the expected values of the outcomes from the two samples [9]:

$$\hat{T}_{mix} = E\left(\frac{\hat{T}_1 * E(y_1) + \hat{T}_2 * E(y_2)}{\left(E(y_1) + E(y_2)\right)} \middle| x_1, x_2, x_3, x_4 \right) \tag{3.1}$$

3.1.1 Blended Methodology II

The blended model, which is the regression of the conditional expectancies of the predictor variables, was developed from the study, where the weights are the variance values of the outcomes from the two samples, and greater weight is given to the result with the lowest variance, as shown below [9]:

$$\hat{T}_{mix,II} = E\left(\left(\frac{\hat{v}_l \hat{T}_s + \hat{v}_s \hat{T}_l}{\hat{v}_l + \hat{v}_s} \right) \middle| x_1, x_2, x_3, x_5 \right) \tag{3.2}$$

3.2 Model-assisted Semi-parametric regression

In this study, model-assisted semi-parametric regression was used to estimate non-parametric regression, which was adapted from [2]. [2] defines a model-assisted estimator as a design-unbiased estimator, as seen below [9]:

$$\hat{T}^* = \sum_s \frac{y_i - m\left(x_j\right)}{\pi_i} + \sum_U m\left(x_j\right) \tag{3.3}$$

While (3.4) is another type of equation (3.3),

$$\hat{T}_{sqrt} = sqrt\left(\sum_{i \varepsilon s} \left(y_i - \beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon \right) * \sum_{i \varepsilon s} \left(y_i - \beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon \right) \right) + \sum_{j \in U \setminus s} \left(\beta_0 + \beta_{x_j} \mathrm{x}_j + \varepsilon \right) \tag{3.4}$$

$$\hat{T}_{3.5} = E\left(\hat{T}_{sqrt} \middle| x_1, x_2, x_3, x_4 \right) \tag{3.5}$$

Also, equation (3.5) is a variant, as is (3.3).

$$\hat{T}_{\pi^*} = sqrt\left(\sum_{i\in s}\left(\frac{y_i-\left(\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)}{\pi_i}\right)*\sum_{i\in s}\left(\frac{y_i-\left(\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)}{\pi_i}\right)\right)+\sum_{j\in U}\left(\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right) \tag{3.6}$$

$$\hat{T}_{3.7} = E\left(\hat{T}_{\pi^*}\middle| x_1, x_2, x_3, x_4\right) \tag{3.7}$$

3.3 The Horvitz-Estimator

In our investigation, the Horvitz-Thompson estimator was used as a baseline to examine the performance of the novel models in terms of coefficient of determination, confidence intervals, sample bias, and standard error [9].

$$\hat{T}_{\pi^*} = \sum_{i\in s_1\cup s_2} w_i^* y_i \tag{3.8}$$

3.4 Model-assisted Semi-parametric Conditional regression

In the study, the Horvitz-Thompson model was tweaked in the same way to develop models that outperformed it. It is the regression of the conditional expectation of the square root of the product of two weighted Horvitz-Thompson estimators [9]:

$$\hat{T}_{\pi^*} = E\left(sqrt\left(\sum_{i\in s_1\cup s_2} w_i^* y_i * \sum_{i\in s_1\cup s_2} w_i^* y_i\right)\middle| x_1, x_2, x_3, x_4\right) \tag{3.9}$$

4. Strategies for combining the two samples

The table 1 below gives strategies used to combined estimates from the two, [9]

[1.] Combining estimates from two samples by blended methodology using expected values of the outcomes		
Blended Methodology I, BMI	$\hat{T}_{mix} = E\left(\frac{\hat{T}_1 * E(y_1)+\hat{T}_2 * E(y_2)}{(E(y_1)+E(y_2))}\middle\| x_1, x_2, x_3, x_4\right)$	
[2.] Combining estimates from two samples by blended methodology using variance values of the outcomes		
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, $\hat{v}_s$= Smaller variance
[3.] Combining estimates from two samples by using residuals		
Semi-parametric regression I, SPRI	$\hat{T} = sqrt\left(\sum_{i\in s}\left(y_i-\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)*\sum_{i\in s}\left(y_i-\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)\right)+\sum_{j\in U\setminus s}\left(\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)$ $\hat{T}_{SPRI} = E\left(\hat{T}\middle\| x_1, x_2, x_3, x_4\right)$	
[4.] Combining estimates from two samples by using weighted residuals		
Semi-parametric regression II, SPRII	$\hat{T}_{SPRII,\pi^*} = sqrt\left(\sum_{i\in s}\left(\frac{y_i-\left(\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)}{\pi_i}\right)*\sum_{i\in s}\left(\frac{y_i-\left(\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)}{\pi_i}\right)\right)+\sum_{j\in U}\left(\beta_0+\beta_{x_j}\mathrm{x}_j+\varepsilon\right)$ $\hat{T}_{SPRII,\pi^*} = E\left(\hat{T}_{\pi^*}\middle\| x_1, x_2, x_3, x_4\right)$	$\pi_i^* = \pi_{1i}+\pi_{2i}-\pi_{1i}\pi_{2i}$
[5.] Combining estimates from two samples by using conditional expectation of the weighted outcomes		
Semi-parametric Conditional Regression, SPCR	$\hat{T}_{SPCR,\pi^*} = E\left(\sum_{i\in s_1\cup s_2} w_i^* y_i\middle\| x_1, x_2, x_3, x_4\right)$	$\pi_i^* = \pi_{1i}+\pi_{2i}-\pi_{1i}\pi_{2i}$
[6.] Combining estimates from two samples by using the Weighted Horvitz-Thompson estimates		
Weighted Horvitz-Thompson, WHT	$\hat{T}_{WHT,\pi^*} = \sum_{i\in s_1\cup s_2} w_i^* y_i$	$w_i^* = \pi_i^{*-1}$ $\pi_i^* = \pi_{1i}+\pi_{2i}-\pi_{1i}\pi_{2i}$

4.1 Strategies used for combining the two samples

Table 2 shows the techniques employed to compare the models.

Strategies used for matching the Models				
s/n	**Model**	**Type of Model**	**Equation of Model**	**Comments**
1.	BM_I	Blended Methodology I	BMI = $\hat{T}_{mix,I}$	Refer to table 1, [1.]
2.	BM_{II}	Blended Methodology II	BMII = $\hat{T}_{mix,II}$	Refer to table 1, [2.]
3.	SPR_I	Semi-Parametric I	SPRI = $\hat{T}_{SPRI}$	Refer to table 1, [3.]
4.*	$SPRII_{\pi^*}$	Semi-Parametric II,π^*	SPRII,π^* = $\hat{T}_{SPRII,\pi^*}$	Refer to table 1, [4.]
5.*	$SPCR_{\pi^*}$	Semi-Parametric Conditional Regression,π^*	SPCR,π^*= $\hat{T}_{SPCR,\pi^*}$	Refer to table 1, [5.]
6.*	WHT_{π^*}	Weighted Horvitz-Thompson	WHT,π^* = $\hat{T}_{WHT,\pi^*}$	Refer to table 1, [6.]

5. Finite small sample properties of estimators

The first attribute is concerned with the mean position of the estimator's distribution.

Biasedness - An estimator's bias is defined as

$$Bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta \tag{5.1}$$

where $\hat{\theta}$ is an estimator of θ, an unknown population parameter. If E ($\hat{\theta}$) = θ then the estimator is unbiased. If E($\hat{\theta}$) $\neq$ θ then the estimator has either a positive or negative bias. That is, the estimator tends to overestimate (or underestimate) the population parameter on average.

A second attribute addresses the variance of the estimator's distribution. Efficiency is a characteristic that is often reserved for unbiased estimators.

Efficiency - Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of θ with equal sample sizes. Then, $\hat{\theta}_1$ is a more efficient estimator than $\hat{\theta}_2$ if

$$\text{var}\left(\hat{\theta}_1\right) < \text{var}\left(\hat{\theta}_2\right) \tag{5.2}$$

6. Results

For each model plan, we ran the simulation 10,000 times. We used software programs built for statistical analysis in R.

Table 3: Summaries of Comparison of Performance for n = 12

s/n	Models	Coeff. Det.	Sample Bias	Standard error	AIC	BIC	Mean	Variance
One Sample								
1.	RC_{One}	0.99686728617276577	0.00089062	0.0020194	42.17611	43.35945	113.7289	697.7151
Two Sample Blending								
2.	BM_I	0.99999999999997935	1.8874e-15	3.0170e-14	-198.6773	-197.4939	113.1562	224.0303
3.	BM_{II}	0.99999999999998446	6.1062e-15	6.9533e-15	-211.3863	-210.203	112.3345	297.6319
Two samples by using residuals								
4.	SPR_I	0.99999999999999600	1.1102e-15	2.9054e-15	-211.7374	-210.554	119.4911	637.9479
Two samples by using weighted residuals								
5.	$SPR_{II\pi^*}$	0.99999999999999756	1.1102e-15	1.2338e-15	-209.9142	-208.7308	117.7882	663.9409
Two samples by using conditional expectation of the weighted outcomes								
6.	$SPCR_{\pi^*}$	0.99999999999999456	2.3315e-15	2.1163e-15	-211.0884	-209.9051	55.72434	236.6314
Two samples by using the Weighted Horvitz-Thompson estimates								
7.	WHT_{π^*}	0.73081861025125028	0.08816741	0.14013330	78.00064	79.18399	55.72434	323.7895

Table 3 shows that the Models have greater coefficients of determination, sample bias and sample standard error than the WHT_{π^*}, and so are more efficient for real Data for n = 12. Where WHT_{π^*} is our reference estimator, the Weighted Horvitz-Thompson Estimator. RC_{one} is the regression calibration done using only one sample and it is evident from the coefficient of determination that ii falls short when compared to the other estimators while it out-performs WHT_{π^*}.

Table 4: Summary of the Performance of Coefficients of Weighted Estimators based on Bias and Standard Error for n = 12

s/n	Estimators	$\hat{\beta}_0$		$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{\beta}_4$	
		Bias	**Error**	**Bias**	**Error**	**Bias**	**Error**	**Bias**	**Error**	**Bias**	**Error**
1	RC_{One}	-98.7971100		-0.0385395		-0.0033699		0.2209781		0.0966125	
		1.87879	57.8639	0.02534	0.51551	0.00067	0.01479	-0.0034	0.09661	-0.0043	0.02808
2	BMI	-22.771935		0.745697		-0.029919		-0.055867		0.167310	
		3.3e-07	9.8e-05	1.9e-07	7.3e-06	-4.1e-09	1.7e-07	-4.8e-08	2.2e-06	2.0e-08	9.3e-07
3	BMII	190.297040		-1.400425		0.026908		-0.182007		0.171984	
		1.4e-06	7.4e-05	-2.3e-08	7.2e-07	-5.0e-10	1.7e-08	-3.3e-09	6.4e-08	4.6e-09	3.2e-08
4	SPR_I	10.556382		-1.068235		0.019519		0.152907		0.110634	
		1.4e-06	4.5e-05	-5.4e-08	3.7e-07	3.7e-10	1.4e-08	3.9e-09	5.6e-08	1.5e-09	3.3e-08
5	$SPR_{II\pi^*}$	-180.532386		0.751331		-0.038316		0.263383		0.058071	
		2.3e-06	3.5e-05	-5.4e-09	3.3e-07	1.0e-09	8.7e-09	2.3e-09	3.4e-08	-6.1e-09	1.9e-08
6	$WHT\pi^*$	-180.532363		0.751331		-0.038316		0.263383		0.058071	
		7.16062	225.862	0.29692	2.32003	0.00316	0.05329	-0.0281	0.23865	-0.0295	0.12627

The results in Table 4 show a summary of the performance of Coefficients of the two sample estimates based on Bias and Standard Error for n = 12, and the summary shows that the coefficients of the two sample estimates appear to have smaller bias and standard errors than the Horvitz-Thompson Estimator for n = 12.

7. Conclusion

Using two sample procedures and regression equations to estimate the coefficients of weighted likelihood regression, this study suggested four approaches for enhancing the efficiency of Regression Calibration. All the five estimators out-performed WHT_{π^*} which was the reference estimator of our study.

8. References

1. Agogo GO. van der Voet H, van 't Veer P, Ferrari P, *et al*. 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. 2014;9(11):1-15. DOI: 10.1371/journal.pone.0113160.
2. Breidt FJ, Opsomer JD. Local polynomial regression estimators in survey sampling, Annals of Statistics. 2000;28(4):1026-1053.
3. Buonaccorsi, JP. Measurement Error: Models, Methods and Application. Chapman Hall/CRC; c2010.
4. Carroll RJ, Ruppert D, Stefanski LA. Measurement Error in Nonlinear Models. Chapman and Hall/CRC; c2006. DOI: https://doi.org/10.1201/9781420010138.
5. Dorfman AH. The two sample problem, Proceedings of the Joint Statistical Meetings, Section of Survey Research Methods. Journal of the American Statistical Association. 2008;87:998-1004.
6. Merkouris T, Combining independent regression estimators from multiple surveys. Journal of the American Statistical Association. 2004;99(468):1131-1139.
7. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Wolters Kluwer Lippincott Williams & Williams; c2008.
8. Thomas D, Stram D, Dwyer J. Exposure Measurement Error: Influence on Exposure-Disease relationships and Methods of correction. Annual Review of Public Health. 1993;14(1):69-93.
9. Kamun SJ, Nyakundi C, Simwa R. Two Sample Approaches to Regression Calibration for Measurement Error Correction. International Journal of Statistical Distributions and Applications. 1993, 2023;9(1):35-40.