Pak. J. Statist.
2023 Vol. 39(4), 457-468

MODIFIED PRINCIPAL COMPONENT CALIBRATION ESTIMATOR IN SURVEY SAMPLING

Muhammad Ahmed Shehzad[1], Haris Khurram[2§], Sharqa Hashmi[3], Adnan Bashir[1] and **Ayesha Niaz[1]**

[1] Department of Statistics, Bahauddin Zakariya University Multan, Pakistan

[2] Department of Sciences and Humanities, National University of Computer and Emerging Sciences, CFD, Pakistan

[3] Govt. Jinnah Associate Collage for Women, Lahore, Pakistan

[§] Corresponding author Email: haris.khurram@nu.edu.pk
ORIC ID: 0000-0003-1814-4742

ABSTRACT

In this article, a modified principal component calibration estimator is proposed by using second moments of principal components for estimating the population total in survey sampling using simple random sampling. A simulation scheme and real-life example are used to evaluate the performance of the proposed estimator. The new estimator is more efficient and reduces bias as compared to the conventional principal component calibration estimator.

KEYWORDS

Survey Sampling, Calibration Estimator, Principal Component Calibration, Moments, Second Order Moments Calibration.

1. INTRODUCTION

Calibration is one of the recently developed techniques in survey sampling literature (see Deville and Särndal, 1992; Singh and Mohl, 1996; Särndal, 2007). The auxiliary variables are used to obtain efficient sample weights called calibration weights such that the weights are the function of auxiliary variables (Deville and Särndal, 1992; Kim and Park, 2010; Farrell and Singh 2005). To obtain the calibration estimation these sample weights are used further. In survey data, the problem of multicollinearity and dimension reduction can be solved by using principal components (PC). Goga and Shehzad (2014), Cardot, Goga and Shehzad, (2017), and Rota and Laitila (2017) used a PC calibration estimator and discuss its properties for high dimensional multicollinear survey data. Bocci and Beaumont (2008) and Goga and Shehzad (2010) discussed the ridge calibration estimator to solve the problem of multicollinearity.

Calibration estimation on different quantiles is discussed by many authors (i.e. Kovacevic, 1997; Harms and Duchesne, 2006; Berger and Munoz, 2015). Using moments of auxiliary variables in calibration estimation is not a very new concept. Kovacevic (1997) used the moment of auxiliary variable in calibration estimation. Moments and their

generalization in calibration estimation are introduced and used by Ren and Deville (2000). Tracy, Singh and Arnab, (2003) used second order moments-based calibration for estimation under stratified random sampling. Harms (2003) used moment-based calibration estimation for small area estimation in the European Household Panel survey. Ren (2002) suggests that calibration on moments is more efficient.

The estimation using dimension reduction and for multicollinear data can be improved after including second moments of principal components (Cardot, Goga and Shehzad, 2017) because it contains information about the variation. Using both first and second moments jointly can be more attractive in achieving high efficiency. By opening the lead provide by Cardot, Goga and Shehzad, (2017) and incorporating Ren and Deville (2000) concept we are in this paper introducing a modified version of PC calibration estimation by adopting the second moment of PCs at the estimation stage. This modification aims to achieve a more efficient estimator of population total when auxiliary variables are of high dimension and multicollinear.

This article is organized as follows. The modification of PC calibration is discussed in Section 2. In Section 3 we discussed the measures used to compare the numerical performance. Section 4 describes the Simulation scheme and its results. For supporting the simulation results Section 5 is developed to perform the said method on a real-life data application. Finally, the conclusion is made in Section 6.

2. MODIFIED PRINCIPAL COMPONENT CALIBRATION

Consider a calibration estimator for population total $t_y = \sum_U y$ is defined by Deville and Särndal, 1992 as:

$$\hat{t}_y = w' y_s. \tag{1}$$

where w is the calibrated weights and y_s is the study variable. The weights w using chi-square distance function is defined as:

$$w = d_s - \prod_s^{-1} X_s (X_s' \prod_s^{-1} X_s)^{-1} (d_s' X_s - 1_U' X)',$$

where d_s is the inverse of inclusion probabilities, $\prod_s = diag(q_{k\in s}^{-1} d_{k\in s}^{-1})$ with q_k as a positive quantity and X_s are auxiliary variables of order $n \times p$ with a property that $w_s' X_s = 1_U' X$.

Assume$Z = (z_1, \ldots, z_r)$ is first r selected principal components (where $r \leq p$). The calibration estimator is now based on these principal components is given by:

$$\hat{t}_{y_{(Z)}} = w_Z' y_s, \tag{2}$$

where w_Z' are the calibrated weights are the function of principal components and defined as:

$$w_Z = d_s - \prod_s^{-1} Z_s (Z_s' \prod_s^{-1} Z_s)^{-1} (d_s' Z_s - 1_U' Z)'.$$

and $w_Z' Z_s = 1_U' Z$.

The variance of the principal components Z is the eigenvalues λ of the components which is the magnitude of features space included and defined as

$$\lambda = \frac{1}{N}\sum_U Z^2.$$

So, adding this supplementary information improves the PC calibration estimator in form of efficiency. Considered, $Z^2 = (z_1^2, \ldots, z_r^2)$ as the second moment of the r principal components. Now instead of using Z in calibrated weights, we used a matrix T, such that $T = (Z, Z^2)$, having order $N \times 2r$. Thus, eq. (2) can be written as:

$$\hat{t}_{y_{(T)}} = \hat{t}_{y_{(Z,Z^2)}} = w_T' y_s, \tag{3}$$

where w_T' are the calibrated weights based on T. Finally, w_T' is defined as:

$$w_T = d_s - \prod_s^{-1} T_s (T_s' \prod_s^{-1} T_s)^{-1} (d_s' T_s - 1_U' T)',$$

$$w_T = d_s - \prod_s^{-1} (Z, Z^2)_s ((Z, Z^2)_s' \prod_s^{-1} T_s)^{-1} \left(d_s' (Z, Z^2)_s - 1_U' (Z, Z^2)\right)'$$

where $w_T' T_s = 1_U' T$ with: $w_T' Z_s = 1_U' Z$ and $w_T Z_s^2 = 1_U' Z^2$.

3. NUMERICAL EVALUATION

To compare the performance of modified PC calibration estimator with PC calibration estimator one we used different measures defined as follows:

- Bias (B): which is measured for an estimated total $\hat{t}_y$ as follows:

$$B(\hat{t}_y) = E(\hat{t}_y) - t_y.$$

- Percent Absolute Bias (PAB): which is measured for an estimated total $\hat{t}_y$ as follows:

$$PAB(\hat{t}_y) = \frac{1}{R}\left|\frac{E(\hat{t}_y) - t_y}{t_y}\right| \times 100$$

- Mean Absolute Error (MAE): which is measured for an estimated total $\hat{t}_y$ as follows:

$$MAE(\hat{t}_y) = \frac{\sum_{i=1}^{R} |\hat{t}_y - t_y|}{R}$$

- Root Mean Square Error (RMSE): which is measured for an estimated total $\hat{t}_y$ as follows:

$$RMSE(\hat{t}_y) = \sqrt{\frac{\sum_{i=1}^{R} (\hat{t}_y - t_y)^2}{R}}.$$

Simulation Study

A Monte Carlo simulation scheme is carried out to compare the performance of the PC calibration estimator with the modified PC calibration estimator. In the data generation process, we have followed a similar scheme as used by Clark and Troskie (2006) and Aslam

(2014) to generate the collinear variables. Ten correlated auxiliary variables have been generated of size 1000 by using the formula:

$$x_{ij} = \sqrt{1-\theta^2}v_{ij} + \theta v_{i,p+1} \text{ where } i = 1,2,\dots,N = 1000; j = 1,2,\dots,p = 10$$

where v_{ij} and $v_{i,p+1}$ are generated as independent standard normal random variates. θ is a controlled value such that θ^2 is the correlation between all auxiliary variables. Four different levels of correlation are considered to check the performance of our estimator between a moderate and high level of multicollinearities by selecting $\theta = 0.80, 0.85, 0.90,$ and 0.95. A study variable y is generated in such a manner that it will depend on the auxiliary variable. To assure their relationship we simply make a linear combination of all auxiliary variables X by adding them and to make this relation inexact we add a small disturbance term u in this model such that

$$y = \sum_{j=1}^{10} X_j + u,$$

where u is independently standard normal distributed variable. Thus, a population total of y denoted by t_y is calculated.

By using PC analysis first five PCs are selected as an auxiliary variable which provides approximately 95% variation of the data. These PCs were used in calibration estimation process. Similarly, for comparison, we have selected 5, 3 and 2 PCs respectively and used these PCs in modified PC calibration estimation process. To check the performance of our estimator for small to large sample sizes, four different samples of size $n = 25, 50, 100,$ and 200 are taken using simple random sampling for each replication. The number of replications R is set to be 1000 for each sample size under each level of multicollinearity. All the data generated including auxiliary variables and study variables are kept fixed in each replication. In last, for each scenario, the estimated total $\hat{t}_y$ is calculated using PC calibration and Modified PC calibration estimator for performance evaluation.

5. RESULTS AND DISCUSSION

Table 1-2 presents the results of different measurements for bias and efficiency of PC calibration and modified PC calibration for different no of PCs at different sample sizes and different levels of multicollinearity. For a small sample size, the B and PAB are lower for the modified PC calibration and there is a considerable reduction in both measurements for the case of modified PC calibration.

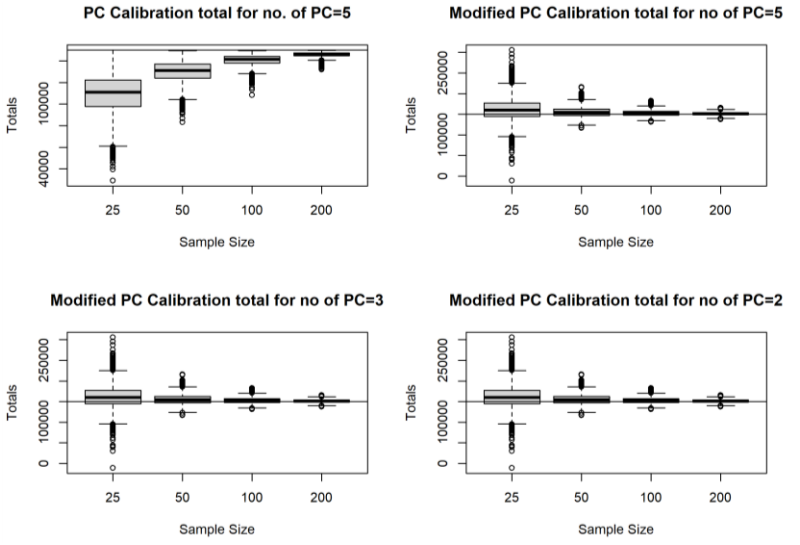
Table 1
Comparison of Bias of PC Calibration and Modified PC Calibration

θ	N	$B\left(\hat{t}_{Y_{Z_{r=5}}}\right)$	$B\left(\hat{t}_{Y_{T_{r=5}}}\right)$	$B\left(\hat{t}_{Y_{T_{r=3}}}\right)$	$B\left(\hat{t}_{Y_{T_{r=2}}}\right)$	$PAB\left(\hat{t}_{Y_{Z_{r=5}}}\right)$	$PAB\left(\hat{t}_{Y_{T_{r=5}}}\right)$	$PAB\left(\hat{t}_{Y_{T_{r=3}}}\right)$	$PAB\left(\hat{t}_{Y_{T_{r=2}}}\right)$
0.80	25	-41097.91	11496.00	11260.52	10805.74	27.39	7.66	7.50	7.20
	50	-20291.96	4968.41	4868.37	4760.21	13.52	3.31	3.24	3.17
	100	-9436.53	2375.49	2297.76	2272.13	6.29	1.58	1.53	1.51
	200	-4242.18	925.46	913.07	903.23	2.83	0.62	0.61	0.60
0.85	25	-40462.29	11260.75	11043.01	10588.53	27.39	7.62	7.47	7.17
	50	-19985.32	4862.10	4766.99	4660.96	13.53	3.29	3.23	3.15
	100	-9294.81	2330.76	2254.28	2229.08	6.29	1.58	1.53	1.51
	200	-4176.36	906.65	894.67	884.75	2.83	0.61	0.61	0.60
0.90	25	-39343.51	10882.96	10690.49	10245.27	27.38	7.57	7.44	7.13
	50	-19440.03	4699.12	4610.00	4507.68	13.53	3.27	3.21	3.14
	100	-9041.94	2258.93	2184.74	2160.30	6.29	1.57	1.52	1.50
	200	-4060.66	877.33	865.90	856.05	2.83	0.61	0.60	0.60
0.95	25	-37329.38	10246.26	10088.76	9666.32	27.38	7.52	7.40	7.09
	50	-18452.62	4429.48	4348.44	4252.45	13.54	3.25	3.19	3.12
	100	-8583.27	2136.24	2066.21	2043.13	6.30	1.57	1.52	1.50
	200	-3852.49	828.22	817.60	808.07	2.83	0.61	0.60	0.59

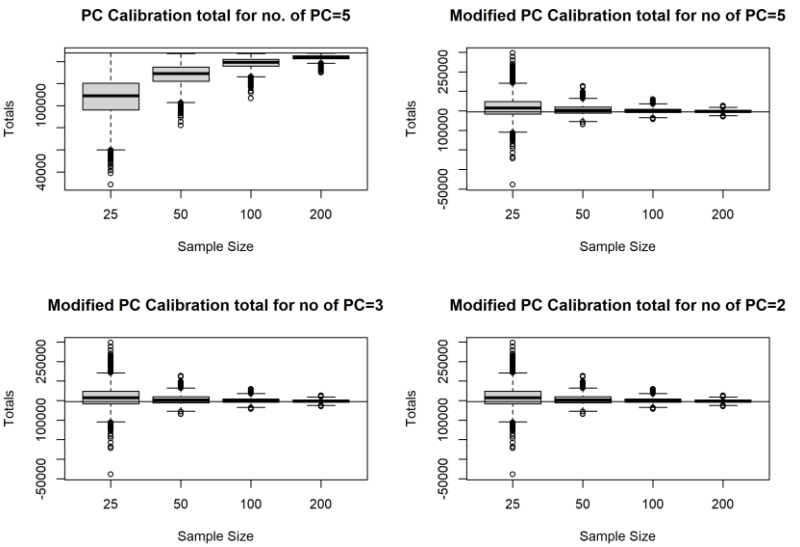
Table 2
Comparison of Efficiency of PC Calibration and Modified PC Calibration

θ	N	$MAE\left(\hat{t}_{Y_{Z_{r=5}}}\right)$	$MAE\left(\hat{t}_{Y_{T_{r=5}}}\right)$	$MAE\left(\hat{t}_{Y_{T_{r=3}}}\right)$	$MAE\left(\hat{t}_{Y_{T_{r=2}}}\right)$	$RMSE\left(\hat{t}_{Y_{Z_{r=5}}}\right)$	$RMSE\left(\hat{t}_{Y_{T_{r=5}}}\right)$	$RMSE\left(\hat{t}_{Y_{T_{r=3}}}\right)$	$RMSE\left(\hat{t}_{Y_{T_{r=2}}}\right)$	$RE = \frac{RMSE(\hat{t}_{Y_{T_{r=5}}})}{RMSE(\hat{t}_{Y_{Z_{r=5}}})}$
0.8	25	41097.91	21798.82	19745.39	18005.90	44812.27	28815.79	26055.60	23719.01	0.643
	50	20291.96	9798.71	9478.02	9141.67	22534.78	12546.99	12204.23	11799.07	0.557
	100	9436.53	5767.28	5659.29	5580.86	10614.66	7363.46	7216.86	7104.96	0.694
	200	4242.18	3349.01	3310.97	3289.30	4806.62	4209.05	4169.90	4138.53	0.876
0.85	25	40462.29	21416.53	19400.09	17684.22	44122.20	28334.91	25600.98	23300.42	0.642
	50	19985.32	9630.42	9315.39	8983.36	22196.64	12323.91	11988.34	11590.31	0.555
	100	9294.81	5664.87	5560.03	5483.02	10454.25	7235.18	7091.88	6980.77	0.692
	200	4176.36	3290.63	3253.45	3231.51	4732.73	4136.90	4098.20	4067.21	0.874
0.90	25	39343.51	20775.42	18830.89	17152.06	42905.81	27522.10	24837.19	22602.88	0.641
	50	19440.03	9346.66	9042.13	8717.51	21593.53	11954.72	11629.94	11243.64	0.554
	100	9041.94	5494.75	5394.75	5320.11	10168.85	7020.61	6882.42	6773.41	0.690
	200	4060.66	3192.93	3156.78	3134.81	4602.40	4015.28	3977.50	3947.28	0.872
0.95	25	37329.38	19650.77	17830.35	16230.75	40713.90	26107.91	23499.91	21387.22	0.641
	50	18452.62	8849.37	8563.01	8253.43	20499.60	11313.27	11006.30	10640.36	0.552
	100	8583.27	5199.23	5106.77	5035.44	9652.01	6645.62	6515.82	6411.36	0.689
	200	3852.49	3022.09	2987.65	2966.33	4367.42	3801.77	3765.78	3737.02	0.870

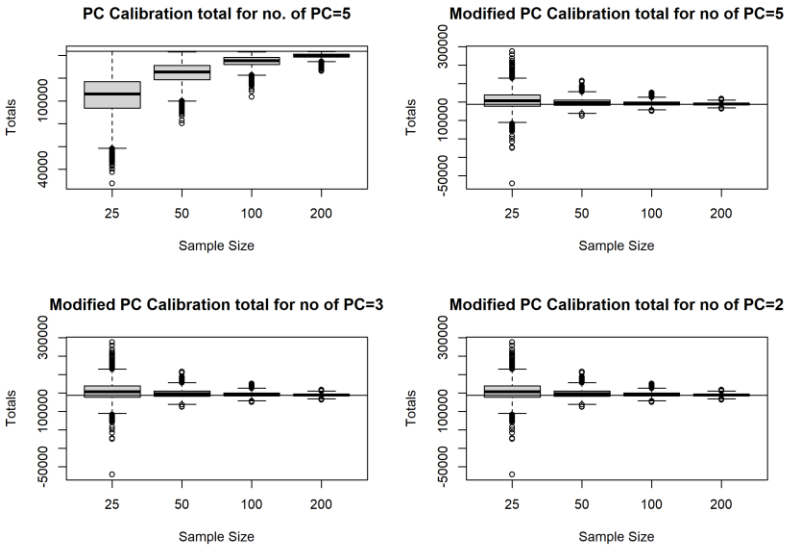
An increase in sample size also causes a considerable reduction in B and PAB when we use modified PC calibration. Similarly, for different no of selected PCs for modified PC calibration, we have a considerable reduction in B and PAB by using less no of PCs as compare to the no of selected PC for conventional PC calibration. The modified methods can perform better when fewer PCs are selected. Moreover, when collinearity among the auxiliary variable becomes severe the B considerably reduces but PAB shows the same results for other levels of multicollinearity.



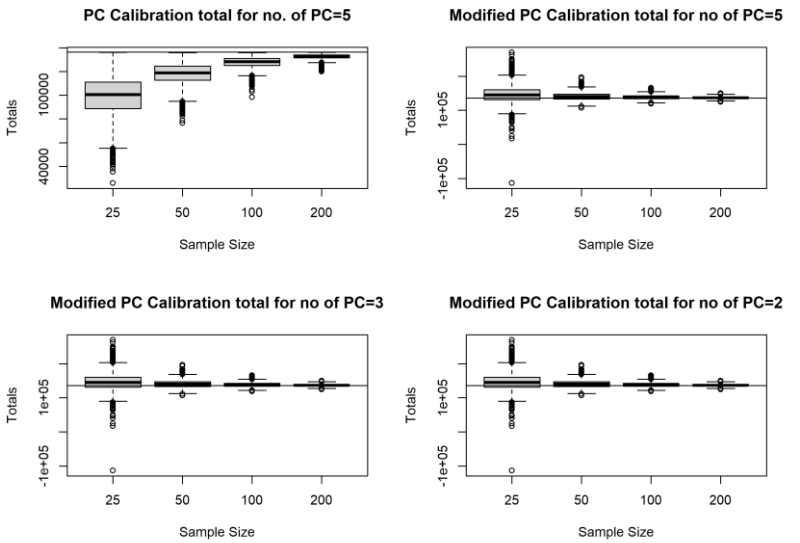
$\theta = 0.80$



$\theta = 0.85$



$\theta = 0.90$



$\theta = 0.95$

Figure 1: Box Plots of the Simulated Total of Modified PC Calibration and PC Calibration

For each sample size, both MAE and RMSE for modified PC calibration are lower than PC calibration. The increment in sample size makes our modified estimator more efficient and it became 26% to 29% more efficient than the conventional estimator. Similarly, for the different numbers of selected PCs, the MAE and RMSE of the modified PC calibration are the lowest than conventional PC calibration.

The same situation can be observed for different levels of multicollinearity. As multicollinearity among auxiliary information becomes high there is an improvement in MAE and RMSE of the modified calibration estimator. Finally, for different no of PCs, at each sample size and level of multicollinearity, our modified estimator is considerably improved in terms of bias and efficiency. Figure 1 shows the boxplot of the simulated totals of simple PC calibration and modified PC calibration with different no. of PCs and different levels of multicollinearity.

5. A REAL DATA EXAMPLE

A real-life example of Boston housing data is also used which was original data by Harrison and Rubinfeld, (1978) and also used by different researches and books. This data was taken from the data archive at http://lib.stat.cmu.edu/datasets/boston consists of 14 variables each has 506 observations. Also available in R with the name "BostonHousing". The study variable used by the different researchers is the median value of owner-occupied homes in USD 1000, so we also take this variable as our study variable (see, i.e. Simlai, 2014; Bargiela, Pedrycz and Nakashima, 2007; Friedman and Wall, 2005; Harrison and Rubinfeld, 1979). The remaining 13 variables are used as auxiliary variables are as follows:

X1: Per capita crime rate by town
X2: Proportion of residential land zoned for lots over 25,000 sq.ft
X3: Proportion of non-retail business acres per town
X4: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
X5: Nitric oxides concentration (parts per 10 million)
X6: Average number of rooms per dwelling
X7: Proportion of owner-occupied units built prior to 1940
X8: Weighted distances to five Boston employment centres
X9: Index of accessibility to radial highways
X10: Full-value property-tax rate per USD 10,000
X11: Pupil-teacher ratio by town
X12: 1000(B - 0.63)^2 where B is the proportion of blacks by town
X13: Percentage of lower status of the population

We are estimating the total median value of owner-occupied homes in USD 1000. This data is a famous data contain high multicollinearity among variables. A small simulation scheme is used to verify the results. For that purpose, we consider the data of 506 observations as our population. After using PCA, seven PCs are selected as auxiliary variables which contain approximately 95% of the total variation. A sample of size $n = 25$, 50, 100, and 200 are take n by using simple random sampling for simulation of size 1000. We use identical variations of PC's, sample size, and simulation size to make a ground comparison between real-life examples and Monto Carlo simulation methods.

Table 3
Comparison of PC Calibration and Modified PC Calibration for Boston Housing Data

N	25	50	100	200
B ($\hat{t}_{y_{(Z)}}$)	-4328.534	-1948.801	-799.257	-278.996
B ($\hat{t}_{y_{(T)}}$)	899.140	278.573	-26.374	1.287
PAB($\hat{t}_{y_{(Z)}}$)	37.964	17.092	7.010	2.447
PAB($\hat{t}_{y_{(T)}}$)	7.886	2.443	0.231	0.011
MAE($\hat{t}_{y_{(Z)}}$)	4332.634	1953.760	812.877	294.529
MAE($\hat{t}_{y_{(T)}}$)	3220.645	1298.045	511.207	263.225
RMSE($\hat{t}_{y_{(Z)}}$)	5042.689	2401.962	998.522	371.040
RMSE($\hat{t}_{y_{(T)}}$)	5011.406	2133.219	673.104	332.453
$RE = \frac{RMSE(\hat{t}_{Y(T)})}{RMSE(\hat{t}_{Y(Z)})}$	0.994	0.888	0.674	0.896

Results are shown in Table 3. The comparison shows a significant reduction in the biases of our modified PC calibration estimator. The measure of bias decreases as the sample size increases. The MAE and RMSE depict that the modified estimator improves its efficiency over simple PC calibration. Figure 2 shows the box plot of the simulated total of both modified and simple PC calibration.

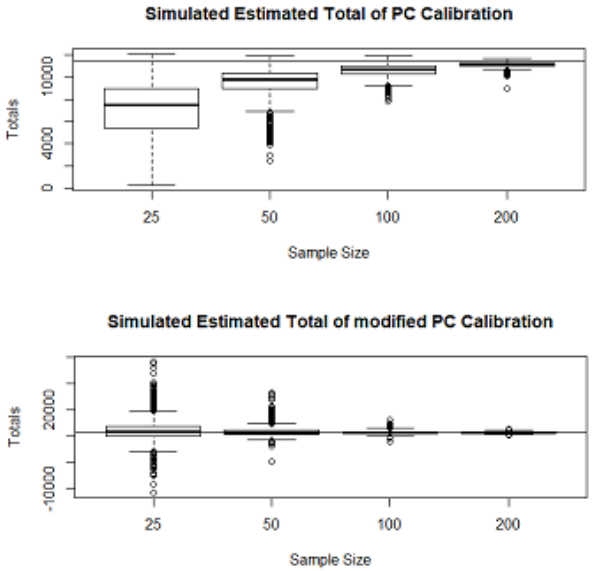


Figure 2: Box Plots of the Simulated Total of Boston Housing Data

6. SUMMARY AND CONCLUSION

This article considers a modification in PC calibration by using the second moment of PCs. The modified PC calibration estimator is compared with PC calibration using bias, PAB, MAE, and RMSE. A simulation scheme and real-life data are used as an example. Results show the significant improvement in bias, PAB, RMSE, and MAE of our modified estimator for different No. of PCs at each sample size and different levels of multicollinearity. It is also concluded that an increase in sample size or levels of multicollinearity shows the better performance of our modified PC calibration estimator over the PC calibration estimator. Moreover using fewer PCs in modified PC calibration also ensure remarkable improvement in the proposed estimator.

DECLARATION

Funding

The authors did not receive support from any organization for the submitted work.

Conflict of Interest/Competing Interest

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

1. Aslam, M. (2014). Performance of Kibria's method for the heteroscedastic ridge regression model: Some Monte Carlo evidence. *Communications in Statistics-Simulation and Computation*, 43(4), 673-686.
2. Bargiela, A., Pedrycz, W. and Nakashima, T. (2007). Multiple regression with fuzzy data. *Fuzzy Sets and Systems*, 158(19), 2169-2188.
3. Berger, Y.G. and Munoz, J.F. (2015). On estimating quantiles using auxiliary information, *Journal of Official Statistics* 31(1), 101-119,
4. Bocci, J. and Beaumont, C. (2008). Another look at ridge calibration. *Metron*, 66(1), 5-20.
5. Cardot, H., Goga, C. and Shehzad, M.A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 243-260.
6. Clark, A.E., Troskie, C.G. (2006). Ridge regression- A simulation study. *Communications in Statistics-Simulation and Computation*, 35, 605-619.
7. Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.
8. Farrell, P.J. and Singh, S. (2005). Model- Assisted Higher- Order Calibration of Estimators of Variance. *Australian & New Zealand Journal of Statistics*, 47(3), 375-383.
9. Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2), 127-136.
10. Goga, C. and Shehzad, M.A. (2010). *Overview of ridge regression estimators in survey sampling*. Université de Bourgogne: Dijon, France.
11. Goga, C. and Shehzad, M.A. (2014). A Note on Partially Penalized Calibration. *Pak. J. Statist.*, 30(4), 429-438.

12. Harms, T. (2003). Extensions of the calibration approach: calibration of distribution functions and its link to small area estimators. *Chintex document de travail*, 13. Federal Statistical Office, Germany.
13. Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey methodology*, 32(1), 37.
14. Harrison Jr, D. and Rubinfeld, D.L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81-102.
15. Kim, J.K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1), 21-39.
16. Kovacevic, M. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. In *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 47, 139-144.
17. Ren, R. and Deville, J.C. (2000). Une généralisation du calage: calage sur les rangs et le calage sur les moments, II čme Colloque Francophone sur les Sondages.
18. Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. Actes d'es Journées de Méthodologie Statistique, *INSEE Méthodes*, Tome 1, 100.
19. Rota, B.J. and Laitila, T. (2017). Calibrating on principal components in the presence of multiple auxiliary variables for non-response adjustment. *South African Statistical Journal*, 51(1), 103-125.
20. Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2), 99-119.
21. Simlai, P. (2014). Estimation of variance of housing prices using spatial conditional heteroskedasticity (SARCH) model with an application to Boston housing price data. *The Quarterly Review of Economics and Finance*, 54(1), 17-30.
22. Singh, A.C. and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-116.
23. Tracy, D.S., Singh, S. and Arnab, R. (2003). Note on calibration in stratified and double sampling. *Survey Methodology*, 29(1), 99-104.