EUROPEAN MASTER IN OFFICIAL STATISTICS
EMOS

Abstract
EMOS Master thesis competition 2025

'Regularizing Probability Sample Estimates Through an Angle-Based Similarity Approach'

Author: **Jacob Westlund, Leiden University**

Keywords: Data integration, Penalized regression, Non-probability samples

1. INTRODUCTION

Today, estimation using probability samples (PSs) is ubiquitous within National Statistical Institutes (NSIs) and using PSs is considered the gold standard due to their strong theoretical properties and relative ease of application. However, collecting PSs is a relatively costly and time-consuming method of acquiring data (Bakker et al., 2014; Van den Brakel, 2019). This has been further exacerbated in recent decades by survey response rates steadily declining, also affecting NSIs who now find it harder to contact and convince individuals to participate in their surveys (De Leeuw & de Heer, 2002; Luiten et al., 2020). For NSIs, such a decline has a twofold impact. First, non-response threatens the estimates' validity by increasing the risk of bias and potentially altering the selection mechanism outside the researcher's control (Bethlehem, 2009). Second, and more practically, non-response also increases the work needed to reach the same sample sizes, which means that either the observed sample sizes will have to be reduced, or more resources need to be allocated to each probability survey sample (Luiten et al., 2020).

Because of this, there has been a growing interest within NSIs in replacing or supplementing the PSs with non-probability samples (NPSs), which are samples obtained outside the probability sampling framework, through for example register data from the tax office or other administrative authorities. This type of register-based NPS is very appealing since it is a very cheap method for acquiring large amounts of data, with no additional burden or requirements for respondents (Van den Brakel, 2019). The hope is that by incorporating these new types of data sources, NSIs can both improve the quality of their estimates and reduce their costs. However, like traditional PS estimates, these new NPS estimates are no silver bullet as they come with their own flaws. Most problematic is the fact that NPSs come without a known sampling designs or inclusion probabilities, and that they tend to suffer from a combination of selectivity and coverage issues. These two factors combined mean that NPS estimates tend to be biased and, lacking a sampling frame, there is no way for NSIs to correct this bias. This is highly problematic for NSIs who place a large value on the unbiasedness of estimates, often preventing NPS estimates from being used directly in official statistics (Bakker et al., 2014).

Therefore, NSIs are currently facing a problem. Both statistically and financially, relying purely on traditional PSs is becoming a prohibitively expensive and troublesome approach. However,

transitioning to the alternative NPSs risks producing biased estimates, which is highly problematic for NSIs whose main purpose is to produce accurate descriptions of a country. Nevertheless, given the high potential of NPSs, the question is raised if there is not some sort of method by which the bias of NPS estimates could be mitigated, directly or indirectly, allowing NPSs to be utilized in official statistics.

Valliant (2020) and Valliant et al. (2018) outline three traditional correction methods, quasi-randomization (also known as propensity score adjustment), superpopulation modelling, and doubly robust estimation. Quasi-randomization is a design-based method where lacking real inclusion probabilities, a PS is used as a surrogate to in some way estimate pseudo-inclusion probabilities for the NPS. Using a common set of auxiliary variables, the PS can be used to estimate the assumed existing but unknown inclusion probability in the NPS. These new pseudo-inclusion probabilities can then be used to re-weight the NPS estimates to reduce or even remove the bias (Elliott & Valliant, 2017; Valliant et al., 2018). Superpopulation modeling breaks with the design-based approach and rather treats the NPS as just a sample from a theoretically infinite "superpopulation", where the outcome of interest follows some unknown probability distribution. The goal of superpopulation modeling is then, given a set of auxiliary variables that explain the selectivity, to use the NPS to model the relationship between the auxiliary variables and estimates of interest. Coefficients can then be extracted and applied to a wider population for population-level statistics that should account for the sample selectivity (Elliott & Valliant, 2017; Valliant et al., 2018). Finally, there is the doubly robust method, which is a combination of quasi-randomization and superpopulation modeling. (see for example Chen et al. (2020)).

These three correction approaches can work, however, they rely on similar and often practically problematic assumptions.

Rather than utilizing the NPS estimates alone, several authors have attempted to harness the information from an NPS to improve the estimates of a related PS instead.

Although the methods differ, the general idea is to maintain approximate unbiasedness of the combined estimates through the PS, whilst leveraging the potentially larger sample size of the NPS to reduce their variances. Elliott and Haviland (2007) and Villalobos-Alíste (2022) looked at integrating an NPS with a PS through a composite estimator. Disogra et al. (2011) proposed an estimation method called "blended calibration" where a calibrated PS is combined with an uncalibrated NPS. The combined sample is then calibrated again using differentiator variables from the PS alone, resulting in a final estimate from the combined sample. Finally, Wiśniowski et al. (2020) applied a Bayesian approach to the incorporation problem, using an NPS to construct priors which are used to estimate a posterior distribution in combination with the PS.

2. OBJECTIVE

The above methods do have their use cases, yet none is designed for producing robust results in a scenario with an unbalanced sample distribution (large NPS and small PS), significant NPS selectivity, and a limited number of auxiliary variables. Correction methods are limited by the lack of auxiliary variables to correct for the NPS selectivity, whilst integration methods on the other hand are more widely applicable, but their results are mixed given very biased NPS estimates and a small PS.

Seeing the larger potential in using integration methods, this thesis seeks to contribute to the literature by proposing an alternative type of integration estimator to the composite, blended

calibration, and Bayesian approach. The estimator draws from the wider literature on penalized regression, which has seen earlier success in integrating estimates from heterogeneous data sources (see Li et al. (2014), Liang et al. (2020), and Tian and Feng (2023)). The general idea of penalized regression for data integration is to use regression with additional penalties. These penalties constrain a set of target estimates of interest towards a set of auxiliary estimates. While we do not care about the auxiliary estimates, they share some similarities with the underlying target estimands. The hope is that this will increase the accuracy of the target estimates by leveraging information about the magnitude and direction of the target estimands of interest from the auxiliary estimates, reducing their variance for only a marginal increase in bias.

Therefore, given its earlier success in general data integration but unknown utility in official statistics, the overarching goal will be to apply a specific type of penalized regression in a context more relevant to NSIs and to answer the question of: How can penalized regression be used to incorporate non-probability samples into official statistics?

3. METHODS

To answer the above-mentioned question, this thesis applies the Angle-Based Transfer Learning Estimator (ABTLE) which is an extension of the traditional ridge regression estimator proposed by Gu et al. (2022). The method is similar to superpopulation modelling and the Bayesian approach of Wiśniowski et al. (2020) in that it is also model-based. However, rather than directly correcting the selectivity, or using the NPS to construct a prior, the ABTLE uses the estimates of the NPS as part of a penalty, seeking to constrain the PS estimates by rewarding them for aligning angle-wise to the estimates of the NPS. This allows for a correction of the PS estimates should the two differ, borrowing the stability of the NPS estimates whilst still through the PS estimates ensuring some protection against the potential bias of the NPS estimates.

The ABTLE assumes a regression model for the vector of target variables y:

$$y=X\beta+\varepsilon$$

where X is matrix with auxiliary data, β is a vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is a vector with normally distributed residuals with mean zero. The vector β can be estimated based on data from PS and on data from NPS. ABTLE and the later mentioned RRE and DBTLE basically start with estimating β using data from PS and adjust this estimate using data from NPS.

The ABTLE has three main advantages over the aforementioned: 1. It is data-cheap, meaning that it only requires estimates directly relevant to the target estimates from the NPS. No actual microdata is required nor are any additional covariates to explain the selectivity of the NPS needed. 2. Given sufficient angle similarity, the degree of bias in the NPS is less impactful on the estimator's quality, making it robust against a very biased NPS, even with a small PS. 3. There is no theoretical risk of negative transfer, meaning the estimate should never be worse than just using the PS estimates in terms of MSE, as long as sufficiently correct penalty parameters are applied.

This thesis applies and evaluates the ABTLE through a simulation study on a wide range of practical scenarios with differing sample sizes, degrees of bias, multicollinearity, sample correlation, and residual variance. This allows for the evaluation of the estimation method in general but also highlights in what context the method is most or least useful. Although the primary metric of

interest is the average root mean squared error (ARMSE), the mean of average bias of the estimator in various scenarios is also of interest given the importance of unbiased estimates in official statistics. The second supplementary goal is to compare this estimation method to other benchmarking estimators such as the Bayesian approach proposed by Wiśniowski et al. (2020) and the ridge regression estimator, to highlight if and when the proposed ABTLE is the appropriate choice.

The ABTLE is compared to standard ridge regression (RRE) and to Distance-Based Transfer Learning Estimator (DBTLE; see Tian and Feng, 2023). Both RRE and DBTLE are special cases of ABTLE, so in principle the results of ABTLE should be at least as good as those of RRE and DBTLE. In practice, this is not always the case due to computational problems when estimating model parameters.

4. RESULTS

The results suggest that the ABTLE can aid in improving the accuracy of ARMSE relative to the PS estimates. The actual improvement however depends a lot on population and sample parameters. Generally, the ABTLE tends to perform better (relatively) in scenarios where the baseline PS estimates were less accurate (higher multicollinearity and larger residual variance).

Regarding the bias, the ABTLE never really outperformed the other estimators, consistently showcasing a higher or merely similar bias as the benchmarking methods. Nevertheless, this is not necessarily problematic since this was most prevalent at smaller levels of bias. As the bias in the NPS estimates increases, the relative bias of all other estimators tends to converge at a really low relative bias to the NPS estimates, showcasing that the ABTLE is also robust when incorporating very biased NPS. The ABTLE only really outperformed the other estimators in scenarios of high correlation but even then the bias was comparable with other methods.

As anticipated, and consistent with the findings of Gu et al. (2022), the results indicate that significant performance improvements in relation to the RRE can be achieved by expanding the penalization scheme to not only penalize magnitudes but also reward aligning it anglewise with auxiliary estimates. Unless the accuracy of the Maximum Likelihood Estimate of the regression coefficient based on the PS only (PS MLE) was already accurate, then even with a flawed penalty parameter estimation this was shown to be true. The results also reveal that a distance-based penalization scheme lacks flexibility as is evident from the fact that the DBTLE tended to underperform compared to the ABTLE. It should be noted that, unlike the results of Gu et al. (2022), it remains unclear from this research whether this issue stems from over or under-penalization of the DBTLE. Finally, relating the results to the research of Wiśniowski et al. (2020), they are in line with their findings. We can see from the relative performance that although the Bayesian estimator works relatively well (sometimes even better than the ABTLE), it is vulnerable to bias in the NPS estimates. As the bias increases to the maximum, the Bayesian approach is never better than the ABTLE, offering further evidence of the limitations of their approach in specific scenarios.

Contrary to the results of Gu et al. (2022), this research also found that the ABTLE can result in negative transfer. Specifically, incorporating the NPS sometimes led to a deterioration in ARMSE relative to the RRE and PS MLE when the accuracy of the PS MLE was already high. This is caused by over-penalization which was not reported as a potential problem in the research of Gu et al. (2022). The likely cause behind both these differences lies in the different simulation designs employed here and by Gu et al. (2022). By chance, it is possible (and was observed) that a sample (or a fold during cross-validation) exhibited a very high correlation or residual variance, inflating the model's

perception of the optimal value of the penalty parameters. This in turn can lead to over-penalization and the occurrences of negative transfer, which does not occur if the theoretically optimal values are used.

Tying the discussion back to the research question of How can penalized regression be used to incorporate non-probability samples into official statistics? the results do highlight a potential type of scenario where the ABTLE does fit. Given its relative success, the ABTLE seems to be a good alternative to the correction methods outlined in Section 2 in scenarios where there are few to no additional covariates to explain the selectivity in the NPS and where the bias of the NPS estimates are large. There is no such thing as the best estimator for every case (Hastie et al., 2009). However, given this type of scenario, correction methods cannot be used, the Bayesian approach and the composite estimator have been shown to be vulnerable to bias. Limited data, high bias scenarios can therefore be seen as a type of scenario where the ABLTE is most likely to be the best choice.

Another more practical scenario where the ABTLE could be a good alternative is in situations where the auxiliary data is sensitive, and sharing micro-level data is not feasible. In such cases, sharing estimates might be more viable which benefits the ABTLE that unlike all other comparable methods only requires auxiliary estimates from the NPS. However, in its current form, there are still clearly some scenarios where ABTLE is to be avoided. With small samples, the estimation of penalty parameters becomes increasingly uncertain, and it is not guaranteed to result in a performance increase. In such scenarios, it might then be preferable to simply use ridge regression since it is more stable, and unlike all other estimators was the only method that never led to deteriorating estimates.

5. CONTRIBUTION

This type of estimation method tends to stem from biostatistics and has to the author's knowledge never been applied in a setting such as the one described above. It is true that one type of estimator proposed by Wiśniowski et al. (2020) is the Bayesian interpretation of the ridge regression estimator (a type of penalized regression), however, it was not a major part of their discussion or research.

6. REFERENCES

[1] Bakker, B. F., Van Rooijen, J. & Van Toor, L. (2014). The System of Social Statistical Datasets of Statistics Netherlands: An Integral Approach to the Production of Register-Based Social Statistics. Statistical Journal of the IAOS, 30 (4), pp. 411–424. https://doi.org/10.3233/SJI-140803.

[2] Bethlehem, J. (2009). Applied Survey Methods: A Statistical Perspective. John Wiley & Sons.

[3] Chen, Y., Li, P. & Wu, C. (2020). Doubly Robust Inference with Nonprobability Survey Samples. Journal of the American Statistical Association, 115 (532), pp. 2011–2021. https://doi.org/10.1080/01621459.2019.1677241.

[4] De Leeuw, E. & De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D. Dillman, J. Eltinge & R. Little (Eds.), Survey Nonresponse, pp. 41–54. Wiley.

[5] Disogra, C., Cobb, C., Chan, E. & Dennis, J. M. (2011). Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. Section on Survey Research Methods – JSM Proceedings, pp. 4501–4515.

[6] Elliott, M. & Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. Survey Methodology, 33 (2), pp. 211–215.

[7] Elliott, M. & Valliant, R. (2017). Inference for Nonprobability Samples. Statistical Science, 32 (2), p. 249–264. https://doi.org/10.1214/16-STS598.

[8] Gu, T., Han, Y. & Duan, R. (2022). Robust Angle-Based Transfer Learning in High Dimensions. https://arxiv.org/abs/2210.12759.

[9] Li, C., Yang, C., Gelernter, J. & Zhao, H. (2014). Improving Genetic Risk Prediction by Leveraging Pleiotropy. Human Genetics, 133 (5), pp. 639–650. https://doi.org/10.1007/s00439-013-1401-5.

[10] Liang, M., Park, J., Lu, Q. & Zhong, X. (2020). Robust and Flexible Learning of a High-Dimensional Classification Rule Using Auxiliary Outcomes. http://arxiv.org/abs/2011.05493.

[11] Tian, Y. & Feng, Y. (2023). Transfer Learning Under High-Dimensional Generalized Linear Models. Journal of the American Statistical Association, 118 (544), pp. 2684–2697. https://doi.org/10.1080/01621459.2022.2071278.

[12] Valliant, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. Journal of Survey Statistics and Methodology, 8 (2), pp. 231–263. https://doi.org/10.1093/jssam/smz003.

[13] Valliant, R., Dever, J. A. & Kreuter, F. (2018). Practical Tools for Designing and Weighting Survey Samples (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-93632-1.

[14] Van den Brakel, J. (2019). New Data Sources and Inference Methods for Statistics, Statistics Netherlands. https://www.cbs.nl/en-gb/background/2019/27/new-data-sourcesand-inference-methods-for-statistics.

[15] Villalobos-Alíste, S. (2022). Combining Probability and Nonprobability Samples on an Aggregated Level, Master thesis, Utrecht University.

[16] Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A. & Blom, A. G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference. Journal of Survey Statistics and Methodology, 8 (1), pp. 120–147. https://doi.org/10.1093/jssam/smz051.