Differentially Private Econometrics through the Analysis of Aggregated Data

Claire McKay Bowen[1], Joshua Snoke[2], Aaron R. Williams[1,*], and Andrés F. Barrientos[3]

[1] *Urban Institute*

[2] *RAND Corporation*

[3] *Florida State University*

[*] *Corresponding author: awilliams@urban.org*

Keywords— administrative data, data confidentiality, data privacy, differential privacy, social science, tax policy analysis

Proposed Article Description: Researchers and practitioners at universities, government agencies, not-for-profits, and industry widely use official statistics and government data to make important public policy decisions, such as how to allocate resources during national emergencies. These data often contain sensitive information and are subjected to various statistical disclosure control methods to protect the confidentiality of individuals or entities in the data. With the introduction of differential privacy (DP) and other formally private definitions (Dwork et al., 2006), the field of statistical disclosure control is experiencing a paradigm shift in the data access and protection landscape (Snoke and Bowen, 2020).

In the U.S., many of the federal statistical agencies are considering or have implemented DP. For example, the U.S. Census Bureau updated their 2020 Disclosure Avoidance System for the 2020 Decennial Census with formal privacy (United States Census Bureau, 2021), and the Internal Revenue Service has recently consider DP and other formally private methods to protect tax data (Barrientos et al., 2021, 2023; Taylor et al., 2021). These changes will affect results derived from official statistics, impact the availability of data, and demand new skills and knowledge for data analysis. They also offer the exciting possibility of releasing new sources of survey and administrative data from the government.

However, Barrientos et al. (2023) and Williams et al. (2023) show there is a large gap between users' expectations of formally private methods and their needs for applied research and econometrics than previously thought. In other words, applied researchers like economists have high expectations for the accuracy of privacy-preserving analysis even when the stakes are very high like sacrificing access to the data (Williams et al., 2023). At the same, DP methods have been shown to return poor results in applied setting for simple econometric analyses like multiple linear regression (Barrientos et al., 2023).

As a follow-up to our previous work, this paper will demonstrate applied research applications where large administrative data and simple summary statistics can support useful applied research with econometrics. Specifically, Barrientos et al. (2023) used differentially private methods that add noise directly to the sufficient statistics of econometric models. In contrast, this paper pursues applications where noise is added while aggregating data (e.g., at the ZCTA level) and then econometric models are applied to the tabulated data. Because DP methods often result in better utility for tabular statistics than regression models, this approach may narrow the gulf between users' expectations of DP and their needs for applied research and econometrics.

In the paper, we will also replicate the research from Feldstein (1995) using DP and a larger sample of administrative tax data. We will pursue at least one more applied example that demonstrates applying statistical methods to DP aggregated data. In all cases, we use Monte Carlo simulation to understand the effects of noise infusion on accuracy and inferences. To evaluate how the results would impact policy decisions, we benchmark our results against estimates of economists' error tolerances gathered from our survey of American Economic Association members (Williams et al., 2023).

We expect the completed article will contribute to the data privacy literature through demonstrations of DP econometrics and the benchmarking of results against the actual expectations of potential users of DP. Our hope is that the results from study will further advance the technical and policy solutions to improving better data access for evidence-based policymaking. Specificially, we will conclude with a discussion of whether or not this approach could benefit a formally private validation server at the National Secure Data Service or IRS.

References

Barrientos, A. F., A. R. Williams, J. Snoke, and C. M. Bowen (2021). Differentially private methods for validation servers.

Barrientos, A. F., A. R. Williams, J. Snoke, and C. M. Bowen (2023). A feasibility study of differentially private summary statistics and regression analyses for administrative tax data. *Journal of American Statistical Association. Accepted*.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pp. 265–84. Springer.

Feldstein, M. (1995, June). The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act. *Journal of Political Economy 103*(3), 551–572.

Snoke, J. and C. M. Bowen (2020). How statisticians should grapple with privacy in a changing data landscape. *Chance 33*(4), 6–13.

Taylor, S., G. MacDonald, K. Ueyama, and C. Bowen (2021). A privacy-preserving validation server prototype.

United States Census Bureau (2021). Disclosure avoidance for the 2020 census: An introduction. Technical report.

Williams, A. R., J. Snoke, C. M. Bowen, and A. F. Barrientos (2023). Disclosing economists' privacy perspectives: A survey of american economic association members on differential privacy and data fitness for use standards. *NBER preprint*.