# Benchmarking Differential Privacy for Applied Public Policy Analysis

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

[1] *Urban Institute*
[2] *Georgetown University*
[3] *Florida State University*
[*] *Corresponding author: awilliams@urban.org*

***Keywords***— administrative data, data confidentiality, data privacy, differential privacy, public policy

**Proposed Article Description:** Researchers and practitioners at universities, government agencies, not-for-profits, and industry widely use government data to make important public policy decisions, such as how to allocate resources during national emergencies and how to evaluate the payoff of attending college. Data that could be useful for answering public policy questions often contain sensitive information and are never made available or are subjected to various statistical disclosure control methods to protect the confidentiality of individuals. Data stewards are using differential privacy (DP) (Dwork et al., 2006) in statistical disclosure control to expand access to administrative data and to potentially support validation servers for interactively running analyses on confidential data. Building on work identifying the needs and expectations of users of administrative data (Williams et al., 2024) and an evaluation of a differential privacy use case (Barrientos et al., 2023), this presentation presents a simulation framework for evaluating differentially private methods for regression. We then evaluate several popular implementations of differentially private linear regression.

In the U.S., some 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 DP (United States Census Bureau, 2021), and the Internal Revenue Service has recently considered DP and other formally private methods to protect tax data (Barrientos et al., 2021). These changes will affect results derived from government data, impact data availability, and demand new analytical skills and knowledge. In order to understand privacy-utility tradeoffs and inform users, work is needed to understand the fitness for purpose of these methods.

Earlier research shows there is a large gap between users' expectations of DP methods and their needs for applied public policy research. In other words, applied researchers 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., 2024). At the same time, DP methods have been shown to

return poor results in applied setting for simple econometric analyses like multiple linear regression (Barrientos et al., 2023).

This paper demonstrates a simulation framework for evaluating the reliability of DP linear regression outputs. The framework is implemented using Monte Carlo simulation and parallel computing to understand the variance of results. The framework benchmarks results against quantitative users' expectations from (Williams et al., 2024) and introduces new metrics based on the variance of results. The framework includes robust tools for introducing violations to assumptions for linear regression and changing the data generating mechanism. Unlike previous work, we can identify the precise scenarios where methods work well on average and where they work poorly.

Using this framework, we evaluate four different differentially private regression methods that support full inference of coefficients. We adjust scenarios by changing sample sizes, the signal-to-noise ratio, the balance of classes for categorical predictors, multicollinearity, the skew of residuals, heteroskedasticity, and omitted variables.

We expect the completed article will contribute to the data privacy literature through demonstrations of DP linear regression methods and the benchmarking of results against the actual expectations of potential users of DP. Our hope is that the results from the study will further advance the technical and policy solutions to improve data access for evidence-based policymaking. We conclude with implications for formally private validation servers at NSDS and IRS, highlighting how benchmarking can guide evidence-based policymaking.

## 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*.

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.

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 (2024). Disclosing economists' privacy perspectives: A survey of american economic association members on differential privacy and data fitness for use standards. *Harvard Data Science Review*.