Differentially Private Linear Regression and Synthetic Data Generation with Statistical Guarantees

In social sciences, where small- to medium-scale datasets are common, linear regression and subsequent statistical inference are central tools for answering scientific questions. In privacy-aware settings, differential privacy (DP) [3] has emerged as the gold standard for providing rigorous privacy guarantees. Various DP methods for linear regression have been proposed, including objective perturbation, DP (stochastic) gradient descent, and sufficient statistics perturbation [1; 7; 8]. However, most existing methods focus on point estimation and risk bounds, offering limited support for uncertainty quantification or being restricted to simple linear regression settings. Only a few works (e.g., [7]) provide procedures for statistical inference, and these rely on restrictive assumptions such as Gaussian covariates and have been shown to yield poor estimation accuracy in practice. As a result, valid statistical inference under DP remains challenging.

A second issue is that reproducibility and replicability are essential to trustworthy social science research [6], but standard DP methods typically release only model estimates, limiting downstream reuse and extension of analyses. While synthetic data generation (SDG) offers a potential remedy, most effective SDG methods with DP protections are tailored to discrete/discretized data and therefore generate discrete outputs (e.g., [5]), whereas deep learning-based approaches (e.g., [4]) preserve data continuity for regression tasks but provide no theoretical guarantees and typically exhibit poor empirical performance as observed in our experiments.

To address these challenges, we propose a novel DP approach for linear regression that simultaneously enables valid statistical inference and synthetic data generation. Our method is built on an efficient and practical binning–aggregation (*BinAgg*) strategy: we use an existing method, PrivTree [9], solely as a building block to obtain a DP partition of the feature domain, then propose a way to aggregate features and labels within bins to reformulate linear regression as a weighted model. Our primary novelty lies in the proposed binning–aggregation framework and the resulting reformulation of the regression model, which supports both valid statistical inference via confidence intervals and the generation of synthetic data. Our approach fills an important gap in the DP literature by supporting *uncertainty-aware linear regression* while simultaneously providing a general mechanism for *regression-aware SDG*. To our knowledge, this is the first work to jointly address DP linear regression with statistically grounded inference and SDG within a single framework, without relying on resampling methods that incur additional privacy costs. Our method demonstrates how multiple analytical tasks can be accomplished under a single privacy budget, achieving both summary-level inference and unit-level reproducibility—two goals that require different data granularities and have previously been addressed only separately.

Main Contributions (1) We propose a novel method, *BinAgg*, for linear regression that satisfies Gaussian DP [2] and achieves one of the best statistical accuracies among existing DP linear regression methods, particularly on real-world datasets with moderate sample sizes and dimensionality. Our method requires minimal tuning and runs significantly faster than competing approaches. (2) We construct valid DP confidence intervals for multivariate settings based on a tailored central limit theorem (CLT), analogous to those in classic non-private regression, without requiring assumptions on the feature distribution. A CLT statement for DP linear regression has been missing from the literature, and our work provides the first such result. (3) Our method also generates DP synthetic data that supports replication studies and downstream tasks beyond linear regression. Empirically, synthetic data generated by *BinAgg* outperforms existing DP synthetic data methods on several downstream machine learning tasks across most real datasets.

References

[1] Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473, 2014.

[2] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

[3] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006.

[4] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

[5] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: an adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.*, 15(11):2599–2612, July 2022.

[6] National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC, 2019.

[7] Or Sheffet. Differentially private ordinary least squares. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3105–3114. PMLR, 2017.

[8] Yu-Xiang Wang. Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[9] Jun Zhang, Xiaokui Xiao, and Xing Xie. Privtree: A differentially private algorithm for hierarchical decompositions. *Proceedings of the 2016 International Conference on Management of Data*, 2016.