Benchmarking DP Linear Regression Methods for Statistical Inference

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

[2] *Urban Institute, awilliams@urban.org & cbowen@urban.org*

[1] *Florida State University, abarrientos@fsu.edu*

[3] *RAND Corporation, jsnoke@rand.org*

Keywords— differential privacy, formal privacy, administrative data, policy analysis, linear regression, econometrics

Abstract: Accessing administrative data is crucial for improving evidence-based policymaking for government officials, policymakers, researchers, and data practitioners. However, direct or secure access to administrative data is often limited to specific government agencies, select researchers collaborating with agency analysts, and highly exclusive research programs. To address these limitations, institutions can employ methods such as synthetic data generation or developing a validation server integrated with differential privacy (DP), to improve data accessibility while ensuring data confidentiality. Initially proposed by Dwork et al. (2006), DP has gained significant traction in recent decades and offers an automated review process, eliminating some elements of human subjectivity. This paper introduces a framework for assessing the performance gap between theoretical expectations and empirical outcomes of DP linear regression methods. Our comprehensive simulation study systematically explores the accuracy and precision of DP regression methods under various scenarios, including heteroscedastic errors and imbalanced categorical variables. We also evaluate the performance of different DP mechanisms, strategies for specifying bounds for unbounded continuous variables, and variations in the data-generating distribution. This simulation study lays the groundwork for assessing future DP regression methods. Finally, we benchmark users' expectations and error tolerances on DP outputs, providing a practical measure for evaluating current DP linear regression methods' viability to meet users' needs.

1 The Value and Limitations of Administrative Data for Economic Research

Accessing administrative data is essential for improving evidence-based policymaking for government officials, policymakers, researchers, and data practitioners. For instance, Nagaraj and Tranchero (2023) demonstrated how direct access to confidential administrative data impacted the rate, direction, and policy relevance of economics research. In particular, one of their findings showed that researchers with confidential administrative data access are more likely to produce papers that receive more citations in public policy documents and produced 24% more publications in top journals per year. Former Under Secretary for Economic Affairs in the Department of Commerce, Jed Kolko, wrote a blog post[1] that states one of the three types of useful research comes from papers focused on "...analyses that directly quantify or simulate policy decisions." Conducting relevant research for impactful policy work often requires accessing administrative data.

Yet, direct or secure data access to administrative data is often restricted to select government agencies, a limited number of researchers working in collaboration with analysts in those agencies, and highly selective research programs run by these agencies. For example, if a researcher wanted access to U.S. taxpayer data, they must be a U.S. citizen before applying for a highly selective research program[2] through the Statistics of Income (SOI) Division at the IRS. If selected, the researcher would then undergo an extensive clearance process that could take several months before gaining access to the data at a secure data enclave. Many researchers in the United States do not meet eligibility requirements, could not be selected for the program, or may not pass the clearance process. Even the use of public use files (PUFs), such as the one produced by SOI, is highly restricted and has become more limited in the amount of information provided over the years due to growing data privacy concerns (Bryant et al., 2014).

To address these issues, multiple reports, such as the Advisory Committee on Data for Evidence Building Year 2 Report[3] and Committee on National Statistics report series on "Toward a 21st

[1] "The economic research policymakers actually need." Accessed April 30, 2024. https://www.slowboring.com/p/the-economic-research-policymakers

[2] "Statistics of Income Joint Statistical Research Program," Accessed April 30, 2024. https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program

[3] "Advisory Committee on Data for Evidence Building." Accessed on April 30, 2024. https://www.bea.gov/evidence

Century National Data Infrastructure[4]," propose creating new tiers of data access. For example, the Urban Institute and SOI are improving the SOI PUF using synthetic data generation (Bowen et al., 2022, 2020, 2022) and developing a validation server (Barrientos et al., 2021, 2024; Taylor et al., 2021) to provide a tier of access between secure data access to the confidential data and the public release of the synthetic SOI PUF.

A well-designed automated validation server system could provide consistent and robust privacy protection with little or no human review, which is both safer and less labor-intensive than current IRS research programs that involve subjective human review. Differential privacy (DP), a concept proposed by Dwork et al. (2006) which has gained substantial traction in the past two decades, provides an attractive option to automate the review process, removing the human element. At a high level, DP and related formal privacy definitions provide an a priori privacy guarantee which when applied consistently enables automatic privacy accounting, under a specific definition.

Several definitions of DP exist, which results in the ability to prove whether a method satisfies that version of DP. We refer to such methods[5] as DP (DP[6]) methods. Satisfying DP is a provable feature of a method, not the data—a common misconception. We provide a brief review of the the mathematical details for DP in the Appendix. While DP offers some desirable properties for expanding access to administrative data, substantial barriers exist to implementing DP in practice for statistical analysis (Snoke et al., 2023).

1.1 The Gap Between Theory and Practice in DP Linear Regression

This paper provides a framework for addressing one of the gaps identified in Snoke et al. (2023), namely the difference between theoretical expectations surrounding the performance of DP linear regression methods and the empirical performance on real data. A substantial literature exists proposing DP mechanisms for linear regression, e.g., Alabi et al. (2022); Barrientos et al. (2024);

[4] "Toward a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century." Accessed on April 30, 2024. https://www.nationalacademies.org/our-work/toward-a-vision-for-a-new-data-infrastructure-for-federal-statistics-and-social-and-economic-research-in-the-21st-century

[5] A note on terminology: In the context of DP, the terms *mechanism*, *algorithm*, and *method* are often used interchangeably to describe the process of releasing a private statistical output. We do not see a clear delineation in the literature when using the three terms. More crucially is that anything referred to as a DP method, DP mechanism, or DP algorithm must provably satisfy the relevant definition of DP.

[6] Note that we use DP as an acronym for both "differential privacy" and "differentially private".

Bernstein and Sheldon (2019); Ferrando et al. (2020, 2021); Sheffet (2017, 2019); Wang et al. (2019); Wang (2018). Some privacy experts have an impression that DP linear regression is a "solved" problem, given the numerous papers and the proposed method's reasonable convergence rates. However, Barrientos et al. (2024) ran the first large scale empirical study of DP multiple linear regression methods on cross-sectional administrative and survey data and found the methods to be largely lacking. The results from these studies showed that very few DP linear regression methods provide multiple coefficient estimates with standard errors, making full inference impossible to conduct. The methods which did produce uncertainty estimates added noise that either inflated the confidence intervals so severely as to limit any conclusions that could be drawn from the data, or the output did not appropriately account for the uncertainty and led to erroneous inferences.

As a follow-up study, Williams et al. (2023) conducted a convenience sample survey of members of the American Economic Association to evaluate the amount of added noise typical economists would tolerate prior to sacrificing access to administrative data. They found that most economists would only accept levels of noise which were far less than the amount added through existing DP linear regression methods at standard privacy budget levels. The conclusions from these studies suggests that current DP linear regression methods can only be utilized at high costs to the accuracy and precision of statistical inference or at high costs to the privacy budget.

We see two main reasons exist for this gap between theoretical expectation and empirical reality. First, we have a finite, often small, sample size when working with real data. This gap has already been recognized as an issue in the DP literature (Slavković and Seeman, 2023) with some methods designed specifically for statistical inference under finite samples in certain cases (Awan and Slavković, 2018; Vu and Slavkovic, 2009). Second, and perhaps more substantially, the simulation studies conducted in papers proposing new DP mechanisms only consider situations where the assumptions of ordinary least squares (OLS) are satisfied and the coefficients have a strong signal. This is not the case in many applications of linear regression for economic, statistical, and social science research. Some examples are the residuals may be skewed or heteroscedastic, there may be multicollinearity between predictor variables, or categorical variables may be imbalanced. We are unaware of any prior work which considers the interaction of adding noise to satisfy DP for

OLS models where one or more of these violations exists.

1.2 A Framework for Benchmarking DP Regression Methods

The contribution of this paper is a framework for explicitly testing DP mechanisms under different scenarios and findings from the application of this framework, so that we can better understand how existing mechanisms will work (or not work) when applied to real data. To create a framework for empirically benchmarking DP linear regression methods under different real-data scenarios, we build on the results from past studies (Barrientos et al., 2024; Williams et al., 2023) to develop a simulation framework that systematically explores the accuracy and precision of performing full inference using the output from DP regression methods for multiple linear regression. Simply put, our simulation study explores circumstances where linear regression assumptions are violated (e.g. heteroscedastic errors) or the estimates without noise are uncertain (e.g., categorical covariates have low-frequency levels). Our findings also help identify the privacy budget level needed under different circumstances to receive sufficiently accurate DP statistics.

We test the best performing DP regression methods identified in Barrientos et al. (2024), which are the Laplace mechanism (Ferrando et al., 2020) and the Analytic Gaussian mechanism (Balle and Wang, 2018). One significant contribution from this paper is providing the infrastructure of our assessment framework for future DP regression methods. We envision two applications for our code base. One is for privacy researchers to assess new DP regression methods, with a focus on statistical inferences rather than predictions. The other scenario involves potential users of a validation server who may wish to formulate an analysis plan based on our assessment framework, similar to a power analysis, on the synthetic data before accessing a validation server that produces formally private outputs. Our code is available online[7] for anyone interested in using this evaluation framework.

Additionally, we use the results from Williams et al. (2023) to benchmark users' expectations and error tolerances on the DP outputs. The purpose of benchmarking against users' expectations and error tolerances is to provide a practical bar against which any viable method must pass. This contributes to the literature by giving a measuring bar for the benchmarking framework that is

[7] GitHub repo website is forthcoming

explicitly tied to users' expectations. We expect future work can and should improve on this tool through repeated surveys of different user groups.

We organize the remainder of the paper as follows. Section 2 outlines how we setup the simulation study, what we consider the baseline scenario, and what assessment metrics we use to evaluate the DP methods. Section 3 evaluates and compares the results across the assessment metrics. We give our concluding remarks and suggestions for future work in Section 4. Details on the definitions, theorems, and mechanisms underlying the DP methods discussed in this paper are provided in the Appendix.

2 Empirical Study Design for Linear Regression under DP

For economists and social scientists, normal linear regression is one of the most popular and commonly used types of regression. In a setting where analysts cannot directly access the data, the goal is to enable them to run a regression analysis while protecting individual private information. Under DP, the analyst receives access to a noisy version of the estimates of interest from the regression analysis. Our empirical study considers the accuracy and precision of these noisy estimates for performing full inference under a variety of different settings.

Various factors can influence a regression analysis, impacting inference and prediction. We group these factors into three categories, with the first two applying to all cases of OLS whether estimated under DP or not. The first group pertains to the data-generating distribution, without implying a violation of the model assumptions. For instance, changes in the signal-to-noise ratio (SNR), probabilities of observing specific categories in categorical variables, correlations among continuous variables, and choices of reference levels in categorical variables. The second group involves violations of model assumptions, such as non-normally distributed errors and non-constant error variance. The last group relates to input parameters necessary for implementing DP, including specifying variable ranges to bound global sensitivity, the privacy budget, and the noise injection mechanism.

We could explore the influence of these factors in a theoretical manner. Although this approach provides much value, we want to avoid making unrealistic assumptions. For example, DP methods

reliance on large sample-based arguments, that lead to the gap between theoretical expectations and empirical reality which we discussed earlier.

We empirically study the influence of these factors through a simulation study. While we acknowledge the limitations of this approach, we argue that empirical studies can shed light on key aspects. They aid in the prioritization of topics for theoretical investigation. Empirical studies can also help practitioners understand which aspects of the data and implementation require caution when employing DP in regression analysis. Furthermore, this empirical study can serve as an example and provide a framework for assessing the influence of such factors when privacy researchers introduce new DP regression approaches.

However, a challenge for studying the influence of the aforementioned factors is that each factor may have many possible settings. The combinations of all the possible levels across each factor results in an enormous number of simulation scenarios. Accordingly, we conduct our empirical study in two stages. In both stages, our main approach is to primarily focus on combinations centered around the levels of one or a few factors, while maintaining the other levels at a reference condition.

In the first stage, we evaluate the performance of the DP mechanisms under *favorable* conditions regarding the data generation mechanism. Specifically, we generate data from an underlying generating distribution that does not violate any of the modeling assumptions. We refer to this scenario as the baseline scenario. The results at this stage help determine the level we fix for some factors in the second stage. For example, we fix the mechanism employed to achieve DP in the second stage after determining the best one in the first stage.

For both stages, we run the DP approaches considering different values of SNR and privacy budget. In the second stage, we assess the performance of the approaches under multiple alternative scenarios characterized by violation of assumptions, multicollinearity, and categorical covariates with low-frequency levels.

After completing the two stages, we leverage findings from both our simulation study and the survey conducted in Williams et al. (2023). Specifically, we delve into economists' perspectives on

DP and usability expectations. Using the tolerance levels identified in Williams et al. (2023), we examine user expectations in detail across the considered scenarios. The results shed light on the implementation of validation servers and provide a guide on the allocation of privacy budgets to meet users' expectations.

2.1 Baseline Scenario

We utilize the following two-stage approach to simulate different linear regression scenarios. We consider the multivariate normal linear regression model.

$$Y = X\beta + e, \tag{1}$$

where $Y = (Y_1, \ldots, Y_n)'$ represents the response and $X = [X_1', \ldots, X_n']'$ defines the design matrix, with $X_i = (1, X_{i,1}, \ldots, X_{i,p})'$. The error, denoted by $e = (e_1, \ldots, e_n)'$, has entries that are independent and identically distributed according to a normal distribution with a mean of zero and a variance equal to σ^2. Here, n denotes the sample size.

We assume that there are $p = 5$ covariates, of which 2 are continuous and 3 are binary, that is, $X_i = (1, X_{i,1} \ldots, X_{i,5})'$.

For the continuous covariates, assume that

$$\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \text{Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{X_1,X_2} \\ \sigma_{X_1,X_2} & \sigma_X^2 \end{pmatrix} \right), \quad i = 1, \ldots, n, \tag{2}$$

where $\sigma_{X_1,X_2} = 0$ and $\sigma_X^2 = 1$.

For the three binary variables, we specifically assume that they represent two categorical variables: one with three categories $\{a, b, c\}$ and another one with two $\{d, e\}$, with probabilities (π_a, π_b, π_c) and (π_d, π_e) such that $\pi_a + \pi_b + \pi_c = 1$ and $\pi_d + \pi_e = 1$. That is,

$$\begin{aligned} \begin{pmatrix} X_{i,3} \\ X_{i,4} \end{pmatrix} &\overset{\text{i.i.d.}}{\sim} \text{Multinomial}(1, (\pi_a, \pi_b, \pi_c)) \\ X_{i,5} &\overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_d), \qquad\qquad i = 1, \ldots, n. \end{aligned} \tag{3}$$

Finally, it is assumed that $(X_{i,1}, X_{i,2})$, $(X_{i,3}, X_{i,4})$, and $X_{i,5}$ are independent.

In the baseline scenario, we simulate $N = 100$ datasets of size $n = 100,000$, setting $\sigma^2 = \sigma_X^2 = 1$, $\pi_a = \pi_b = 1/3$, and $\pi_d = 1/2$. The choice of a sample size of 100,000 is primarily influenced by Barrientos et al. (2024), which conducted regression analysis under DP using datasets of a similar scale. This study serves as a follow-up to that work, hence the consideration of this specific sample size. Additionally, theoretical assumptions in DP linear regression often necessitate sufficiently large datasets to ensure certain properties. Future work will explore smaller sample sizes to better understand the impact of sample size on accuracy and precision of statistical estimates. Different values for β are also considered to achieve various SNR, defined as

$$\text{SNR} = \frac{\text{Var}(X_i\beta)}{\sigma^2}. \tag{4}$$

The β values are set such that $\text{SNR} = \{0, 0.5, 1, 3, 6\}$, resulting in a total of 500 simulated datasets.

To perform regression analysis under DP, we privatize $S = [X, Y]^t[X, Y]$. This term is the sufficient statistic for the normal linear model, resulting in the DP statistic $S_H = S + H$, where H represents the added noise. We incorporated the noise using the ϵ-DP Laplace mechanism (Dwork et al., 2006) and the (ϵ, δ)-DP Analytic Gaussian mechanism (Balle and Wang, 2018). We employ two different strategies to generate inferences about the regression coefficients β. The first strategy relies on a modified version of Algorithm 3 from Ferrando et al. (2020). We use Monte Carlo simulation to approximate the sampling distribution of an estimator of β based on S_H, accounting for the randomness in both e and H. The modified version of their algorithm can be found in Algorithm 1 in the supplemental material of Barrientos et al. (2024). It introduces regularization to ensure that S_H is positive definite and addresses the challenge of having an unknown sample size n. The second strategy employs the regularized version of S_H described in Barrientos et al. (2024) and substitutes it into the formulas of the maximum likelihood estimator without DP. This strategy is justified by Theorem 1 in Ferrando et al. (2020), which demonstrates that the effect of H becomes negligible when the sample size is sufficiently large, particularly regarding the asymptotic distribution of the estimator of the regression coefficients. In testing all our methods, we run the DP methods using $\epsilon = \{0.5, 1, 5, 10, 10^6\}$ and $\delta = 10^{-7}$.

For both the Laplace and Analytic Gaussian mechanisms, it is necessary to specify lower and upper bounds for the response variable and the two continuous covariates. Specifying the bounds of these variables is crucial when implementing DP because the larger the distance between the lower and upper bounds, the larger the variance of noise produced from the mechanism. Additionally, there is a trade-off. Narrower bounds decrease the variance of the mechanism but may introduce some bias in the estimates, and vice versa. For this reason, we consider three different strategies to set these bounds.

The first strategy sets the bounds for a given dataset using the observed minimum and maximum, which is equivalent to using local sensitivity or the perfect information scenario. We refer to these bounds as local bounds. The second strategy uses the maximum and minimum of each variable over the simulated 100 datasets for a given SNR, and we refer to them as group bounds. The third strategy defines the bounds using a DP mechanism for quantiles proposed by Gillenwater et al. (2021). The idea is to use part of the privacy budget to query large and small quantiles and use them as bounds. The queried quantiles we use are those that accumulate 0.001 and 0.999 of the relative frequency. Although the approach proposed by Gillenwater et al. (2021) also requires specifying bounds as input parameters, we decide to specify those bounds to be roughly 100 times wider than observed ones. We believe that providing bounds that are excessively large is a more realistic situation for an analyst compared to expecting the use of bounds that closely match the observed range.

2.2 Assessment Metrics

To assess the effect of satisfying DP in linear regression, we rely on metrics that compare coefficient estimates, uncertainty measures, and predictions to those produced under no privacy and to the underlying truth. For each scenario, SNR, and scenario-specific condition, we simulate $N = 20$ datasets of size $n = 100,000$. The employed metrics are described below.

- *Effective sample size.* In Bayesian statistics, a prior distribution for the parameters in the normal linear model is the unit information prior (see Hoff, 2009, for a description). This prior distribution is derived from the fact that the Fisher information based on a sample of n observations is $\sigma^{-2}X^TX$ and the "average information" in one observation is $\sigma^{-2}X^TX/n$. Since

$\sigma^2(X^TX)^{-1}$ is the covariance matrix of the maximum likelihood estimator of the regression coefficients $\hat{\beta}$, we can consequently argue that the "average variance" of the j-th regression coefficient estimator resulting from using a single observation is given by $n\text{Var}(\hat{\beta}_j)$, where $\text{Var}(\hat{\beta}_j)$ is the j-th component of the diagonal of $\sigma^2(X^TX)^{-1}$.

Under DP, there is also the variance associated with the regression coefficient estimates $\hat{\beta}^{\text{dp}}$. We denote the variance of the j-th regression coefficient estimator under DP as $\text{Var}(\hat{\beta}_j^{\text{dp}})$. We can then define the effective sample size $n_{\text{ESS},j}$ as the number of observations such that a variance without privacy would equal $\text{Var}(\hat{\beta}_j^{\text{dp}})$. Therefore, $n_{\text{ESS},j}$ is of the form

$$n_{\text{ESS},j} = n\frac{\text{Var}(\hat{\beta}_j)}{\text{Var}(\hat{\beta}_j^{\text{dp}})},$$

which can be estimated using estimates of $\text{Var}(\hat{\beta}_j)$ and $\text{Var}(\hat{\beta}_j^{\text{dp}})$.

- *Coverage probability and bias.* For both private and nonprivate approaches, we compute point estimates and confidence intervals for each β_j. Subsequently, we employ Monte Carlo simulation to approximate the absolute and relative bias of the point estimators and the coverage of the confidence intervals. For the coverage probability, we compute the fraction of times that the interval contains the true value of β_j. Ideally, this probability is expected to be close to the specified confidence level.
- *Matching sign and significance.* We check if the sign of $\hat{\beta}_j^{\text{dp}}$ matches that of $\hat{\beta}_j$, and if the decision on $H_0 : \beta_j = 0$ versus $H_A : \beta_j \neq 0$ remains consistent with and without privacy. We utilize Monte Carlo simulation to approximate the probability of matching signs and decisions.
- *Confidence interval ratio.* We calculate a ratio of the confidence intervals with noise to the confidence intervals without noise, which gives us a measure of the increased uncertainty due to the added noise.

2.3 Alternative Scenarios

Based on the results of the baseline scenario analyses, we proceed to run each alternative scenario using only one of the DP mechanisms while still considering different values of SNR, different privacy budgets, and different bounds. We consider the following alternative scenarios.

- *Imbalanced categorical predictor.* For this scenario, we simulate data considering different values of π_d in Equation 3. The values for π_d range from 0.01 to 0.99. When the values for π_d are large, it implies that the values for π_e are small. Under this scenario, we can observe the effect of having a dummy variable where the probability of being equal to one is low, as well as the effect of having a reference level with low frequency, which is expected to impact the intercept.
- *Collinearity.* In this setup, we simulate data where the continuous predictors are correlated to varying degrees, that is, considering different values of σ_{X_1,X_2} in Equation 2, ranging from moderate to high levels of correlation. We consider $\sigma_{X_1,X_2} \in \{0, 0.5, 0.9\}$ This scenario allows us to observe how collinearity affects the estimation of regression coefficients and the associated uncertainties under DP.
- *Heteroscedasticity.* The normal linear model assumes that the errors follow a normal distribution with constant variance. Heteroscedasticity violates this assumption. We can simulate the data in different ways that violate this assumption. We choose to simulate data such that the variance of the errors σ^2 in model 1 is a function of the continuous variable X_1. Specifically, we replace σ^2 in model 1 with $\sigma^2_{X_1} = (1-\tau/2)+\tau\exp(X_1)/(1+\exp(X_1))$, where $\tau \in [0, 2]$ controls how fast $\sigma^2_{X_1}$ grows as a function of X_1. If $\tau = 0$, then $\sigma^2_{X_1} = 1$ for any value of X_1. If τ increases, then $\sigma^2_{X_1}$ increases at a faster rate. Figure 1 displays $\sigma^2_{X_1}$ for the values of τ we consider $\{0, 0.5, 1, 2\}$, representing no, low, medium, and high heteroscedasticity, respectively. In this scenario, a scatter plot of X_1 versus e would display a pattern often observed in practice, resembling a fan or cone shape. The definition of $\sigma^2_{X_1}$ ensures that $E(\sigma^2_{X_1}) = 1$. Under this scenario, we determine the value of β assuming that the SNR is redefined as $\text{Var}(X_i\beta)/E(\sigma^2_{X_1})$.
- *Skewed Residuals.* Another assumption we consider is the normality of the errors. In this scenario, we test the effect of having errors that are not normally distributed but instead asymmetric. To do so, we rely on the skew normal distribution. A random variable Z_i is skew-normal distributed if its density is defined as $2\phi(z)\Phi(\alpha z)$, where α is a parameter which regulates asymmetry, and ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively. To ensure that the

errors e_i have zero mean and variance equal to one, we define them to be equal to $(Z_i - E(Z_i))/\sqrt{Var(Z_i)}$. Figure 1 displays the different degrees of asymmetry we consider by setting α to take values in $\{0, 1, 3, 50\}$, representing no, low, medium, and high asymmetry, respectively.

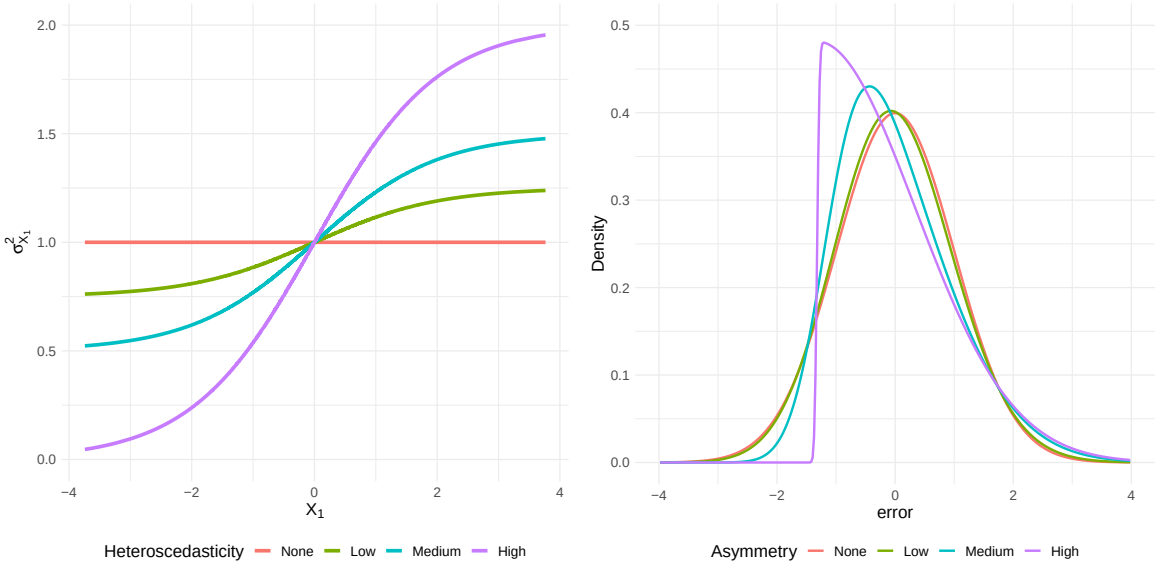

Figure 1: Two scenarios considered in the simulation study: heteroscedasticity and asymmetry. The left plot illustrates varying levels of heteroscedasticity, where the variance of errors $\sigma^2_{X_1}$ is controlled by the parameter τ, which takes values in $\{0, 0.5, 1, 2\}$ representing no, low, medium, and high heteroscedasticity, respectively. The right plot displays the distribution of errors generated using the skew normal distribution with parameter α regulating asymmetry. As α varies across $\{0, 1, 3, 50\}$, different degrees of asymmetry are represented, ranging from none to high.

2.4 Users' Expectations

Finally, we compare the results of the simulations to the findings in Williams et al. (2023) based on a survey of American Economic Association members' views on DP and the usability of noise-infused data. One aspect of the survey focused on querying respondents about their tolerance for error before sacrificing access to administrative data. Specifically, the questionnaire asked about four different ways that error might affect the results of a statistical query: (1) significance mismatch of confidential and noisy statistics, (2) sign mismatch of confidential and noisy statistics, (3) absolute relative bias in the point estimate, and (4) the confidence interval ratio between confidential and noisy confidence intervals.

Using the tolerances quantified in Williams et al. (2023), we use this simulation study to

investigate whether users' expectations are met under the baseline and alternative scenarios, as well as different values of the privacy budget. This additional analysis provides practical values against which we assess the benchmarking simulation. For example, it helps determine how much of the privacy budget might need to be allocated to ensure that the released results meet expectations.

3 Results

3.1 Baseline Simulations

Table 1 shows the relative bias for each mechanism and each value of the privacy-loss budget (ϵ). Relative bias is the estimated bias of the estimate divided by the population parameter. All of the relative biases are less than 0.01, which indicates a bias of less than 1%. The relative biases are comparable for the Laplace mechanism and Analytic Gaussian mechanism.

	Mechanism	ϵ	Simulated repetitions	Relative bias
1	analytic gaussian	0.5	2400	-0.0052
2	analytic gaussian	1	2400	-0.0003
3	analytic gaussian	5	2400	-0.0018
4	analytic gaussian	10	2400	-0.0016
5	analytic gaussian	1,000,000	2400	-0.0008
6	laplace	0.5	2400	-0.0032
7	laplace	1	2400	-0.0030
8	laplace	5	2400	-0.0016
9	laplace	10	2400	-0.0015
10	laplace	1,000,000	2400	-0.0008

Table 1: Summary Relative Bias for Mechanism and Privacy-Loss Budget. The table only shows results using the bootstrap confidence intervals and DP range method for determining sensitivity. The estimated statistics combine different signal-to-noise ratios and coefficients results. Simulated repetitions is the number of coefficients for each relative bias estimate

Table 2 shows the coverage probability for a 95% confidence interval for each mechanism and each value of the privacy-loss budget (ϵ). All of the values of the coverage probability are close to 95%. This holds true for small values of ϵ, which suggests the confidence intervals account for the large amount of noise added to the regression models by the DP mechanisms. The coverage probabilities for the Laplace mechanism and Analytic Gaussian mechanism are comparable.

Figure 2 shows the effective sample sizes for each estimated coefficient using the Laplace and Analytic Gaussian mechanisms with sensitivities determined by the observed ranges of the data.

	Mechanism	ϵ	Simulated repetitions	Coverage probability
1	analytic gaussian	0.5	2400	0.9450
2	analytic gaussian	1	2400	0.9542
3	analytic gaussian	5	2400	0.9492
4	analytic gaussian	10	2400	0.9429
5	analytic gaussian	1,000,000	2400	0.9396
6	laplace	0.5	2400	0.9446
7	laplace	1	2400	0.9592
8	laplace	5	2400	0.9483
9	laplace	10	2400	0.9379
10	laplace	1,000,000	2400	0.9400

Table 2: Summary Coverage Probability for a 95% Confidence Interval for Mechanism and Privacy-Loss Budget. The table shows results using the bootstrap confidence intervals and DP range method for determining sensitivity. The estimated statistics combine different signal-to-noise ratios and coefficients results. Simulated repetitions is the number of coefficients for each coverage probability estimate.

The results are limited to three coefficients and the bootstrap method of estimating standard errors. The ideal effective sample size, 100,000 observations, is represented by the horizontal red line.

First, Figure 2 shows the dramatic reduction in effective sample size caused by the DP mechanism even for large values of ϵ. For example, when $\epsilon = 5$ and the SNR is 3, the effective sample size of the intercept approximately drops by one third and the effective sample size of the continuous variable X_1 drops by a half. Second, the figure shows the reduction in sample size worsens as the SNR increases. Finally, the figure demonstrates that the two mechanisms have comparable results. We will focus on the Laplace mechanism because it has a slightly stricter privacy guarantee and does not require specifying δ in addition to ϵ.

Figure 3 compares the effective sample size of estimates using the local bounds with the alternative approaches of using the DP range bounds and grouped local bounds. The diagonal red line indicates equivalent effective sample size for each comparison. The points are overwhelmingly above the red line for the DP range bounds. Extreme percentiles (0.001, 0.999) are approximations of minima and maxima and introduce truncation error. In the baseline scenario, the introduced truncation error is smaller than the reduction in error caused by using smaller sensitivities when fitting the noisy regression model. However, the DP range bounds use the exponential mechanism, which provides extremely poor results in a small fraction of cases. The differences between the local

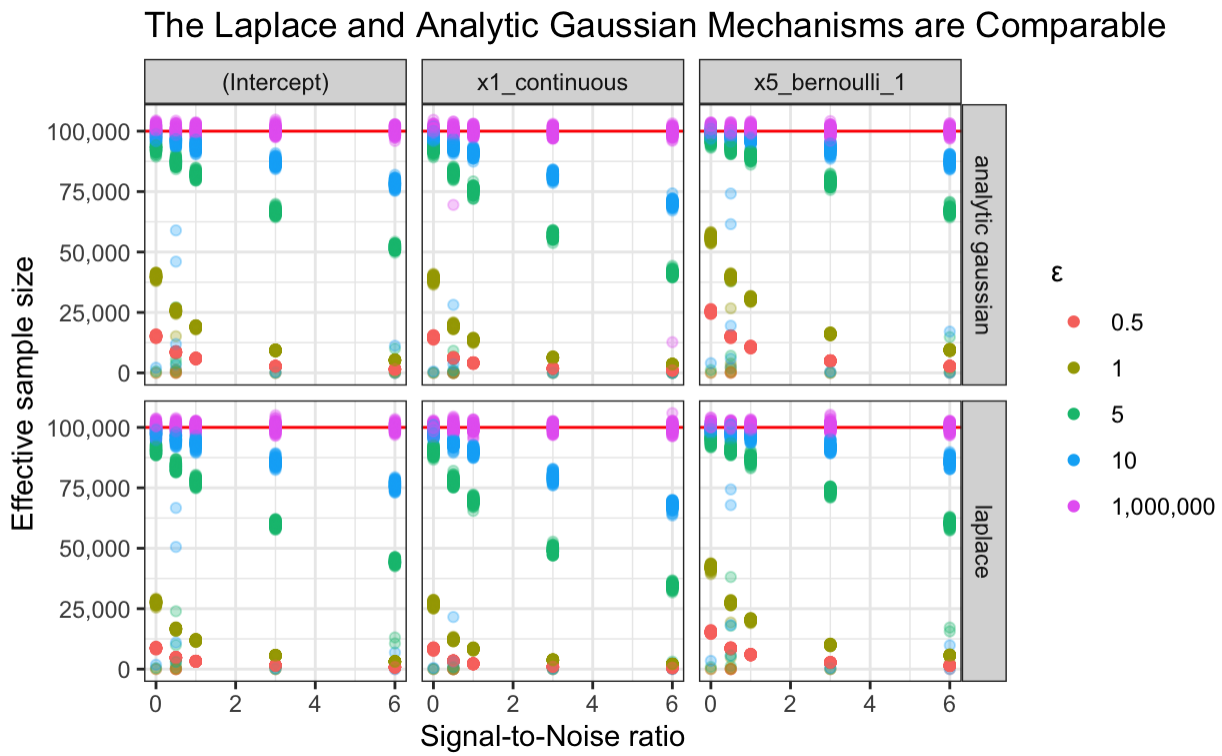

Figure 2: Figure includes the results for selected coefficients and are limited to the bootstrap confidence intervals and DP range bounds.

bounds and DP range bounds are dramatic. When $\epsilon = 5$ and the SNR is 3, the effective sample size of the intercept drops by a half instead of dropping by one third. The effective sample size of the continuous variable X_1 drops by two thirds instead of halving when using the local bounds instead of the DP range bounds. The grouped local bounds consistently lowers the effective sample size of estimates more than the local bounds. This makes sense because the grouped local bounds is always greater than or equal to the local bounds.

Absent clear ranges on the data, for example, limiting ages to prime working ages, calculating sensitivities with DP percentiles is a realistic process that could be used on an automated validation server. Unless otherwise noted, we will focus exclusively on determining the sensitivities using the DP range approach (DP range bounds).

Figure 4 shows the coverage probabilities for asymptotic and bootstrap confidence intervals for different values of ϵ and different methods for calculating sensitivities. The asymptotic confidence intervals have poor coverage probabilities for most values of ϵ. Alternatively, the bootstrap

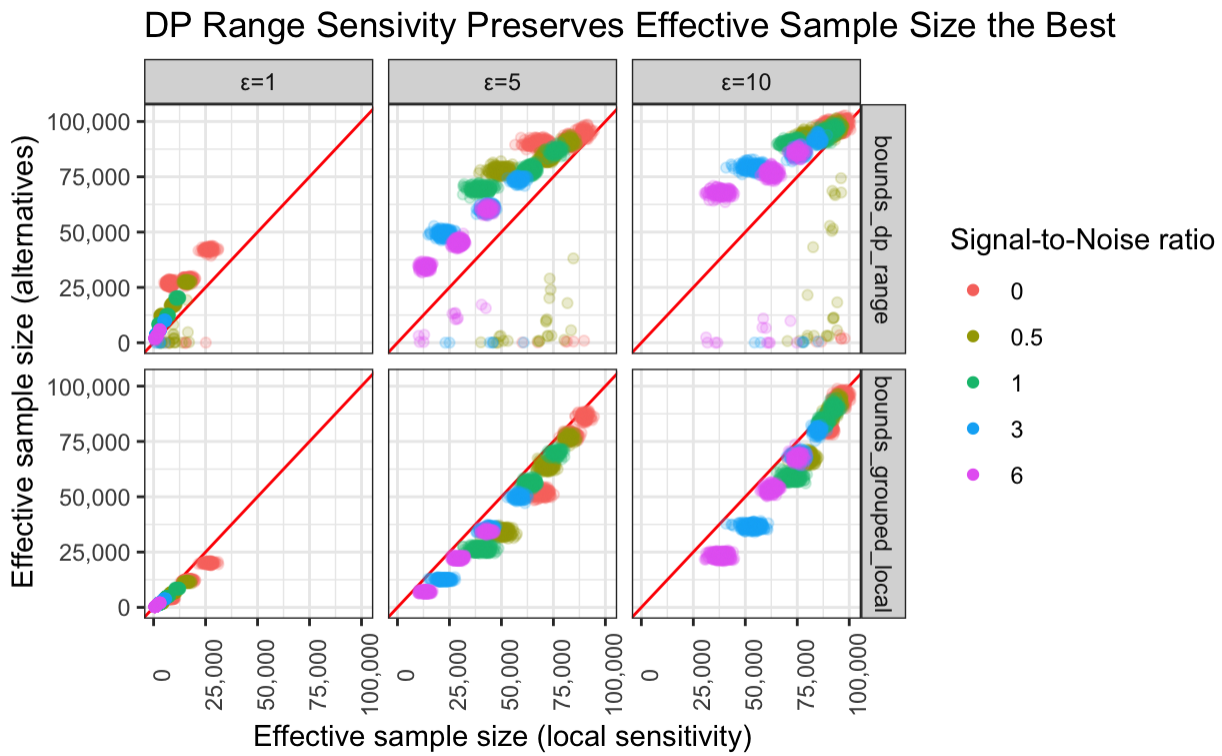

Figure 3: Figure shows $\epsilon \in \{1, 5, 10\}$ and uses the Laplace mechanism and bootstrap confidence intervals.

confidence intervals demonstrate reliable coverage probabilities for all three sensitivities. For this reason, we focus on the bootstrap confidence intervals for all results.

3.2 Alternative Scenario Simulations

Table 3 shows the three worst relative biases for each detailed scenario and each value of the privacy-loss budget (ϵ). The table combines SNR and coefficients to boost sample sizes. All of the relative biases are less than 0.01, which indicates a bias of less than 1%.

Table 4 shows the coverage probability for a 95% confidence interval for each detailed scenario $\epsilon = 5$. The table combines SNR and coefficients to boost sample sizes. Differences can largely be explained by the number of data replicates for the alternative scenario. The baseline run uses 100 data replicates and nearly perfectly hits 0.95. The values of the alternative scenarios are slightly noisier because they rely on 20 data replicates.

Table 3 suggests that all alternative scenarios have reasonable amounts of bias. Table 4 suggests that all alternative scenarios have reasonable coverage probabilities. Next, we look at the effective

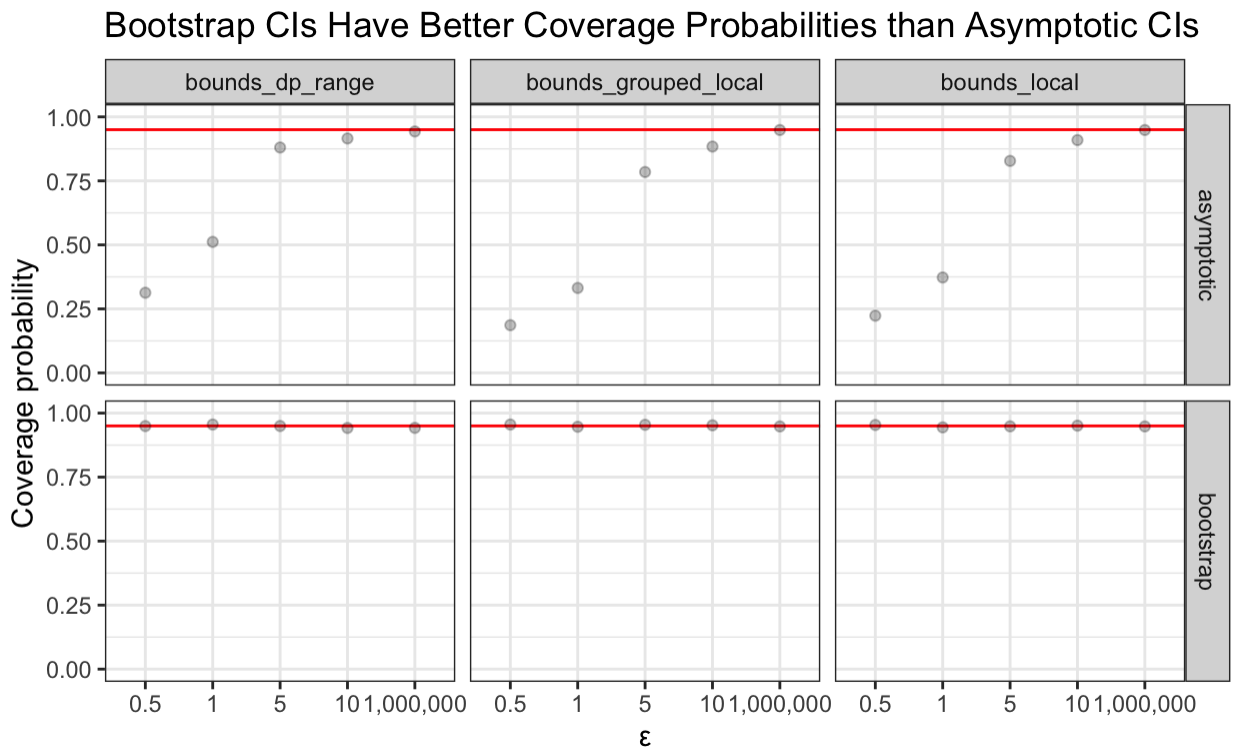

Figure 4: The results are a mix of coefficients, terms, and signal-to-noise ratios. Each point represents 3,000 coefficients.

sample sizes for each alternative scenario for $\epsilon = 5$. We choose to present the results for this privacy budget value because it was the smallest one identified in Barrientos et al. (2024), which yielded reasonable performance for DP approaches most of the time.

3.2.1 Class Imbalance

Figure 5 shows the effective sample size for the intercept, one of the continuous variables, and the Bernoulli variable when $\epsilon = 5$ for different SNR and different amounts of class imbalance. When $\pi_d = 0.5$, the probability for the Bernoulli random variable is 0.5 and the scenario aligns with the baseline scenario. As π_d increases, the effective sample sizes for the intercept and X_5 diminish. As π_d decreases, the effective sample size for X_5 diminishes. The effective sample sizes for other estimates in the model are mostly unchanged.

Including categorical variables with sparse categories as predictors in linear regression models does not violate any assumptions of linear regression models, but it does increases variances for estimates of the coefficient for the categorical variables. Here, it appears that DP may add an extra

	Deatiled scenario	ϵ	Simulated repetitions	Relative bias
1	imbalanced categories 0.5	0.5	480	0.0044
2	multicollinearity 0.9	0.5	480	-0.0040
3	imbalanced categories 0.1	0.5	480	0.0035
4	imbalanced categories 0.01	1	480	0.0051
5	baseline	1	2400	-0.0030
6	skewed residuals 3	1	480	-0.0024
7	baseline	5	2400	-0.0016
8	imbalanced categories 0.01	5	480	-0.0015
9	heteroscedasticity 2	5	480	-0.0015
10	heteroscedasticity 2	10	480	-0.0015
11	baseline	10	2400	-0.0015
12	skewed residuals 50	10	480	-0.0013
13	heteroscedasticity 2	1,000,000	480	-0.0014
14	skewed residuals 50	1,000,000	480	-0.0012
15	imbalanced categories 0.99	1,000,000	480	-0.0012

Table 3: Three Worst Biases for Detailed Scenarios and Privacy-Loss Budgets. Results only include the Laplace mechanism with sensitivities determined with DP ranges and bootstrap confidence intervals. Simulations with Signal-to-Noise Ratio = 0 are excluded. Simulated repetitions is the number of coefficients for each relative bias estimate.

penalty to estimates of coefficients in the presence of sparse categories.

3.2.2 Multicollinearity

Figure 6 shows the effective sample size for the intercept, both continuous variables, and the Bernoulli variable when $\epsilon = 5$ for different SNR and different amounts of covariance between the two continuous predictors. When covariance σ_{X_1,X_2} is 0, the scenario aligns with the baseline scenario. As covariance increases, the effective sample size for X_1 and X_2, the two continuous variables decreases. The effective sample sizes for other estimates in the model are mostly unchanged.

Like with class imbalance, including predictors with non-perfect collinearity does not violate any assumptions of linear regression models. It does, however, increase the variance of estimates of the coefficients for the collinear predictors. Here, it appears that DP may add an extra penalty to estimates of coefficients in the presence of collinearity.

	Deatiled scenario	ϵ	Simulated repetitions	Coverage probability
1	baseline	5	3000	0.9497
2	heteroscedasticity 0	5	600	0.9467
3	heteroscedasticity 0.5	5	600	0.9417
4	heteroscedasticity 1	5	600	0.9200
5	heteroscedasticity 2	5	600	0.9450
6	imbalanced categories 0.01	5	600	0.9550
7	imbalanced categories 0.1	5	600	0.9567
8	imbalanced categories 0.5	5	600	0.9550
9	imbalanced categories 0.9	5	600	0.9633
10	imbalanced categories 0.99	5	600	0.9400
11	multicollinearity 0	5	600	0.9617
12	multicollinearity 0.5	5	600	0.9300
13	multicollinearity 0.9	5	600	0.9317
14	skewed residuals 0	5	600	0.9367
15	skewed residuals 1	5	600	0.9467
16	skewed residuals 3	5	600	0.9467
17	skewed residuals 50	5	600	0.9350

Table 4: Summary Coverage Probability for a 95% Confidence Interval. The table shows the results to the Laplace mechanism, bootstrap confidence intervals, and DP range bounds. Simulated repetitions is the number of coefficients for each relative bias estimate.

3.2.3 Skewed Residuals

Figure 7 shows the effective sample size for the intercept, one continuous variable, and the Bernoulli variable when $\epsilon = 5$ for different amounts of skewness in the residuals and different approaches to determining sensitivities. When $\alpha = 0$, the scenario aligns with the baseline scenario. The effective sample sizes don't change when α is changed.

Errors that aren't normally distributed are typically only a problem for linear regression models with small sample sizes. DP dramatically reduces effective sample sizes. But, with 100,000 observations, the effective sample sizes in these simulations are typically large enough to avoid issues with skewed residuals. It's possible that smaller sample sizes would have worse results, but these simulations don't show any specific challenges for skewed residuals.

The different rows in figure 7 demonstrate the effect of different bounds on the effective sample sizes. The middle row shows the grouped local bounds. This means using the minima and maxima within a SNR for the bounds of the data. This plausible approach dramatically reduces the effective

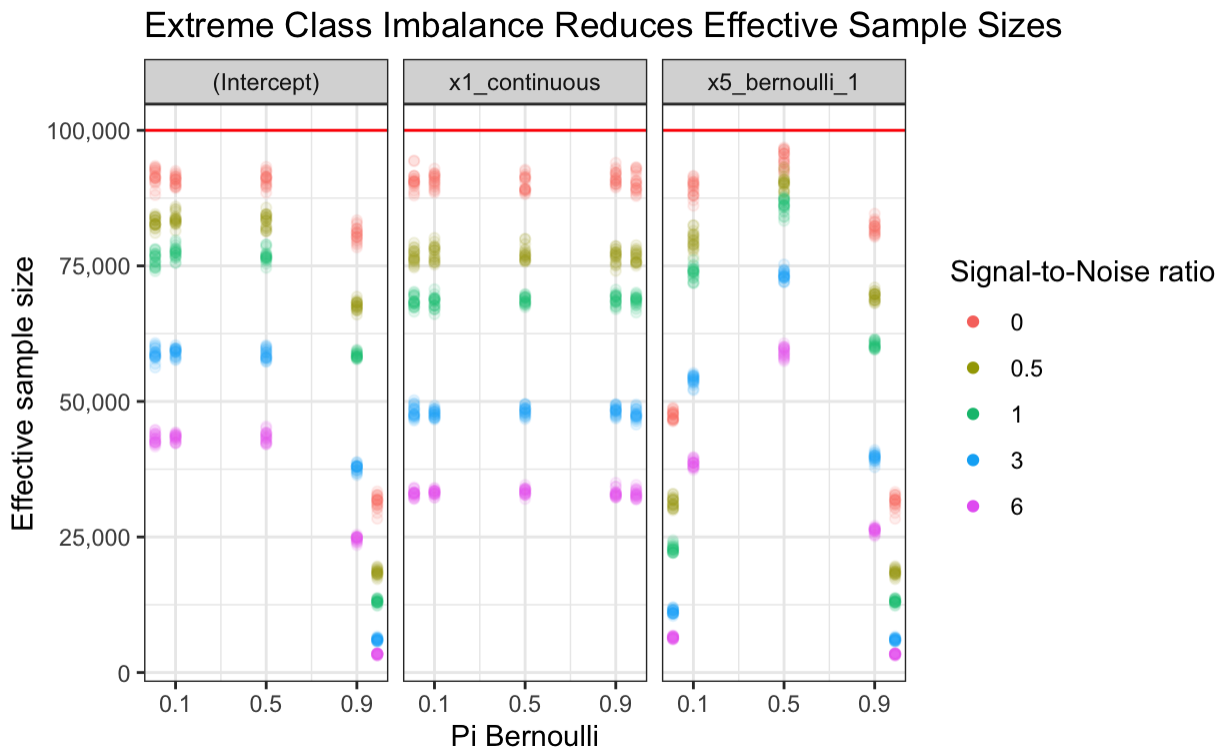

Figure 5: Figure includes results for selected coefficients and are limited to $\epsilon = 5$, the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

samples compared to the local bounds and DP range bounds.

3.2.4 Heteroscedasticity

Figure 8 shows the effective sample size for the intercept, one continuous variable, and the Bernoulli variable when $\epsilon = 5$ for different amounts of heteroscedasticity and different approaches to determining sensitivies. When heteroscedasticity is 0, the scenario aligns with the baseline scenario. The effective sample sizes remain constant when heteroscedasticity is changed.

Heteroscedasticity limits the ability to generalize variance estimates without necessarily increasing variance estimates. The average variance of the residuals remains unchanged in the alternative scenarios even though the shape of the residuals changes. So the effective sample sizes remain constant. It's possible that a different pattern for heteroskedasticity would result in worse results. It's also likely that the results have poorly calibrated prediction intervals and could result in poor inferences in certain scenarios.

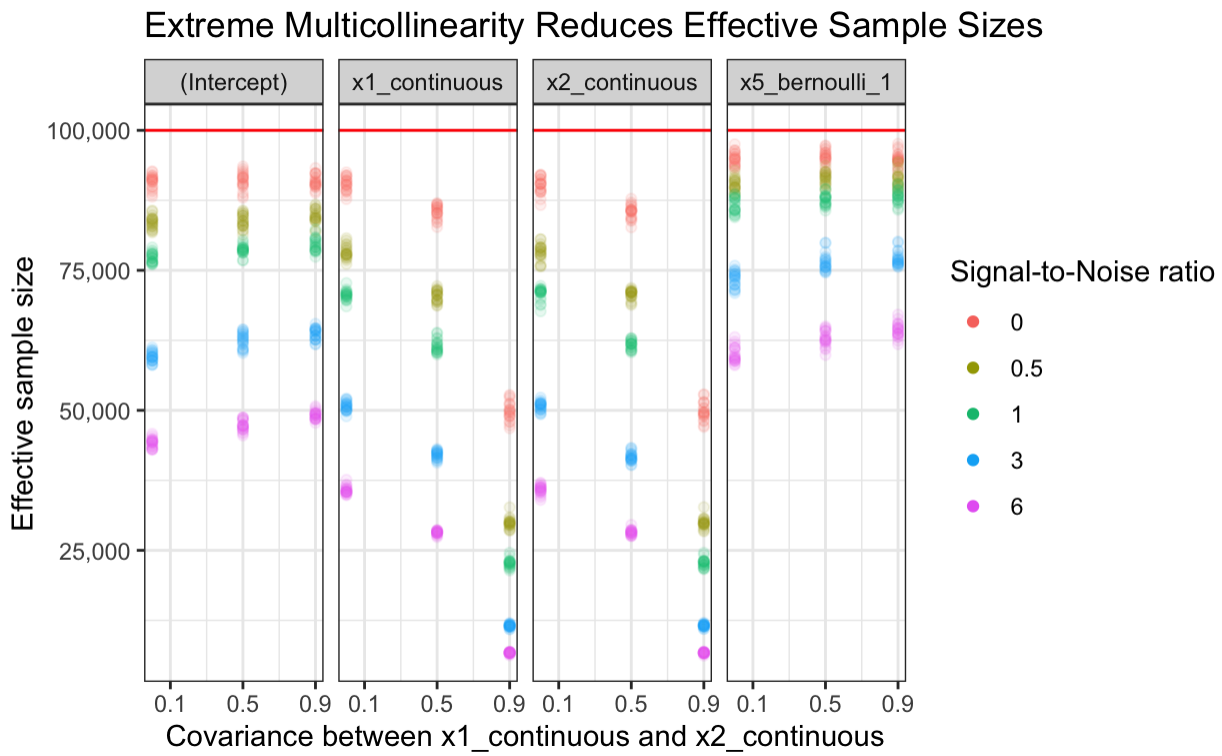

Figure 6: Figure includes results for selected coefficients and are limited to $\epsilon = 5$, the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

The different rows in figure 8 demonstrate the effect of different bounds on the effective sample sizes. The middle row shows the grouped local bounds. This means using the minima and maxima within a SNR for the bounds of the data. This plausible approach dramatically reduces the effective samples compared to the local bounds and DP range bounds.

3.3 Users' Expectations

Effective sample size, bias, and coverage ratio are useful ways to evaluate DP regression methods. In this section, we use a different set of metrics to evaluate the noisy regression output against users' expectations from Williams et al. (2023). Their paper surveys members of the American Economic Association and asks respondents how much error they would tolerate before sacrificing access to the administrative data. Here, we focus on sign mismatch, significance mismatch, absolute relative error, and confidence interval ratio.

Sign mismatch is the relative frequency with which a noisy estimate is expected to have a different sign (positive or negative) than an estimate without noise (Williams et al., 2023). The

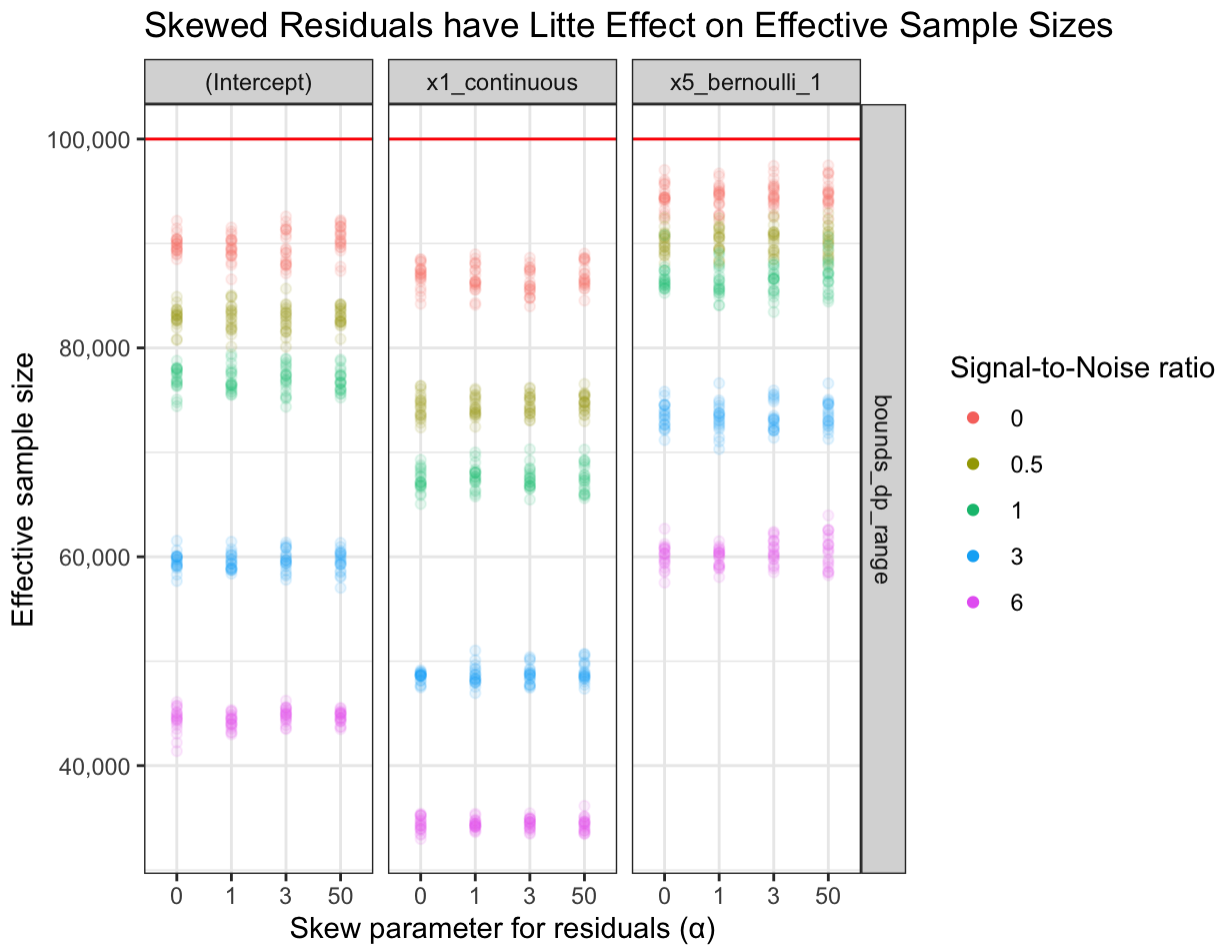

Figure 7: Figure includes results for selected coefficients and are limited to $\epsilon = 5$, the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

median respondent would tolerate a sign mismatch rate of 5%. Figure 9 shows the proportion of simulations that meet the median user's expectations for each coefficient, ϵ, and SNR and for the most extreme parameterization of each alternative scenario. Many simulations have sign mismatch when the SNR is 0. The overwhelming majority of baseline simulations and alternative scenario simulations with a positive SNR have no sign mismatch. The one exception is class imbalance, although the failure to meet users' expectations disappears with larger values of ϵ.

Significance mismatch is the relative frequency with which a noisy estimate has a different statistical significance (assume 0.05 level) than the estimate without noise (Williams et al., 2023). The median respondent would tolerate a significance mismatch rate of 10%. Figure 10 shows the proportion of simulations that meet the median user's expectations for each coefficient, ϵ, and SNR

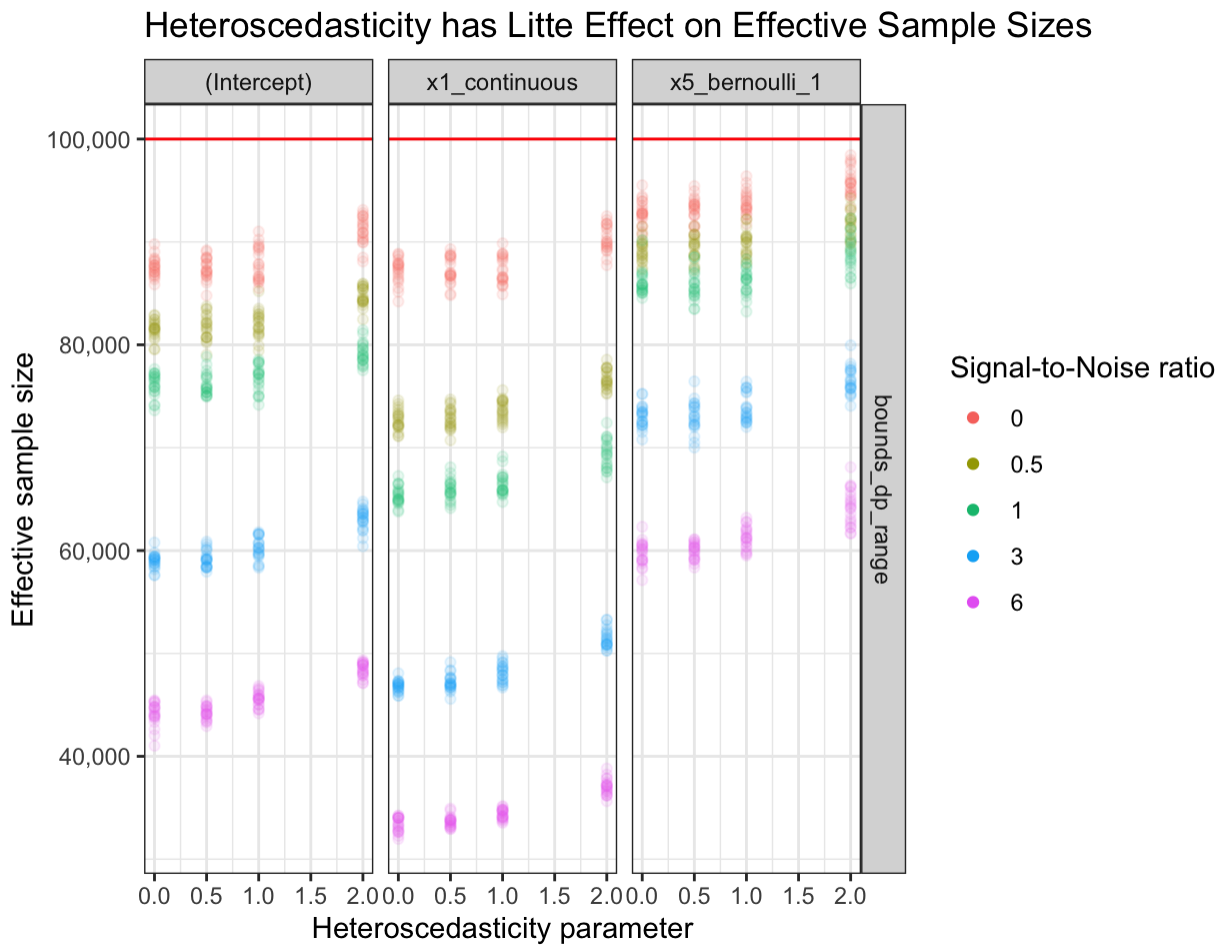

Figure 8: Figure includes results for selected coefficients and are limited to $\epsilon = 5$, the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

and for the most extreme parameterization of each alternative scenario. Many simulations lead to significance mismatch when the SNR is 0. The overwhelming majority of baseline simulations and alternative scenario simulations with a positive SNR have no significance mismatch. The one exception is class imbalance, although the failure to meet users' expectations disappears with larger values of ϵ.

Absolute relative error is the amount of noise introduced into an estimate divided by the estimate without noise (Williams et al., 2023). The median respondent would tolerate absolute relative error of 10%. Figure 11 shows the proportion of simulations that meet the median user's expectations for each coefficient, ϵ, and SNR and for the most extreme parameterization of each alternative scenario.

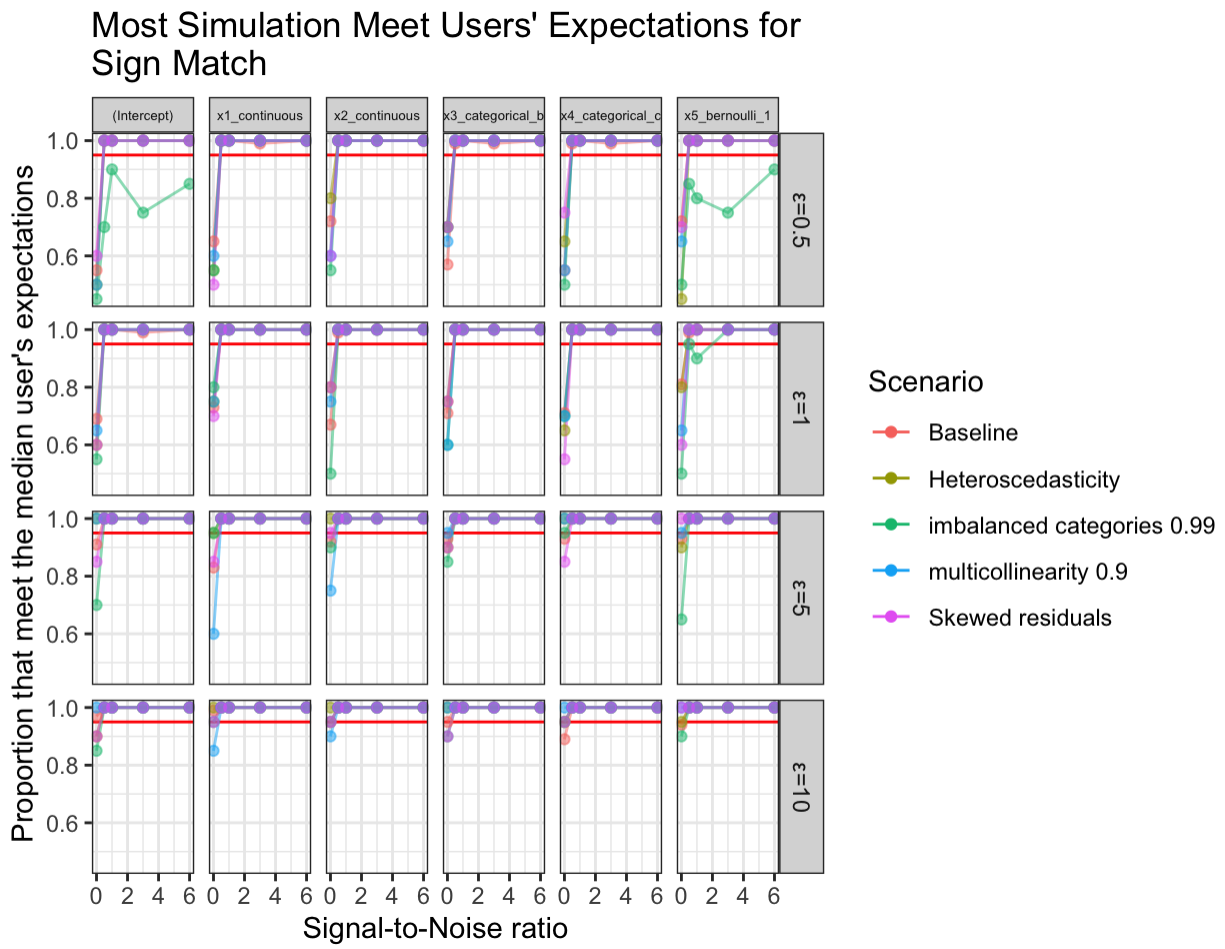

Figure 9: Each line represents the most extreme specification for a scenario. The results are limited to the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

Many simulations lead to large absolute relative errors when the SNR is 0. This is likely because the estimate without noise, which is in the denominator of the metric, is exceptionally close to zero. The overwhelming majority of baseline simulations and alternative scenario simulations with a positive SNR meet the median user's expectations. Specific coefficients fall short for the class imbalance and multicollinearity scenarios, whereas the failure to meet users' expectations disappears with larger values of ϵ.

Confidence interval ratio is the width of the noisy confidence interval divided by the width of the confidence interval without noise (Williams et al., 2023). The median respondent would tolerate a confidence interval ratio of 1.25. That means the noisy confidence interval is 25% wider than the confidence interval without noise. Figure 12 shows the proportion of simulations that

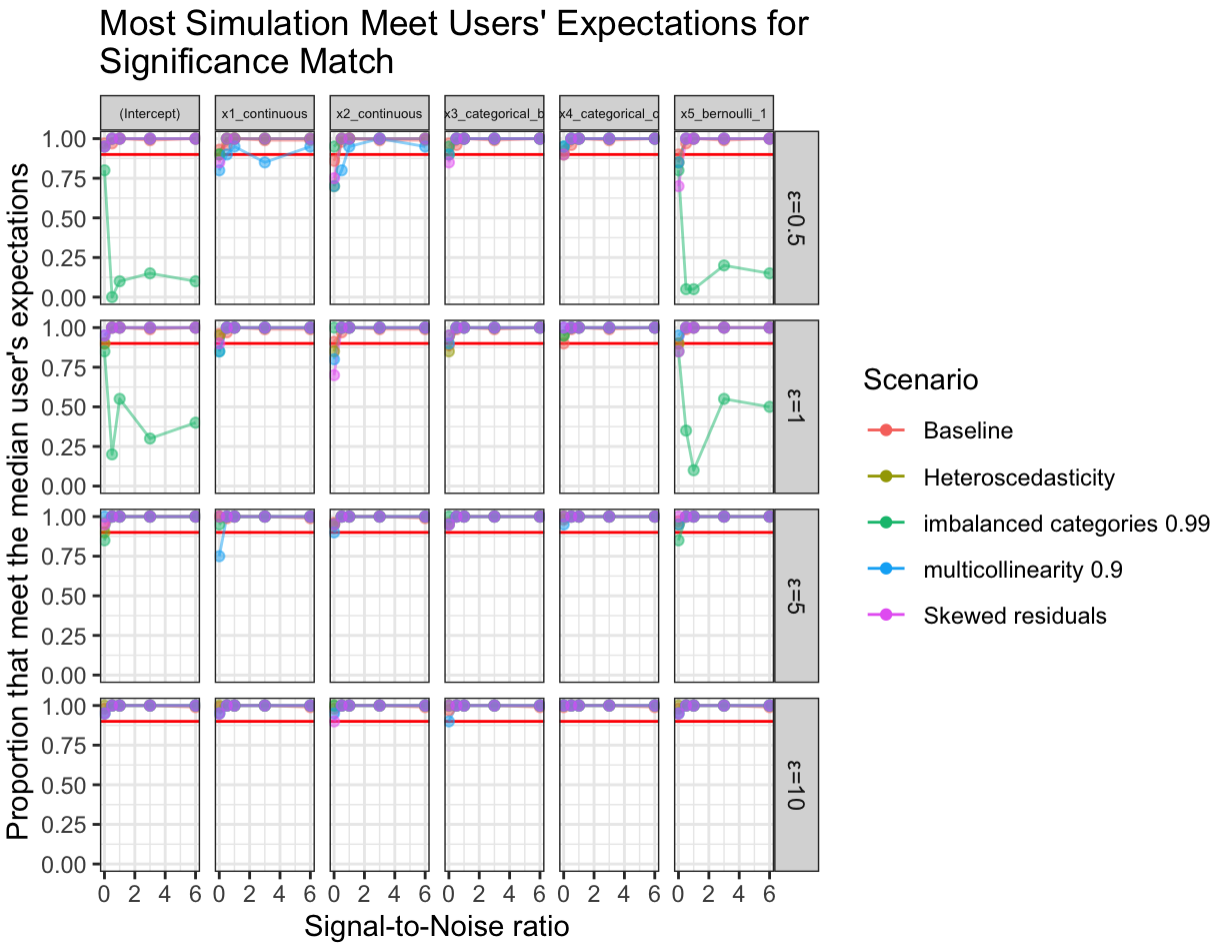

Figure 10: Each line represents the most extreme specification for a scenario. The results are limited to the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

meet the median user's expectations for each coefficient, ϵ, and SNR and for the most extreme parameterization of each alternative scenario.

None of the simulations meet the median user's expectations when $\epsilon < 5$. When $\epsilon = 5$, the width of the confidence interval varies with the SNR and the scenario. Even when $\epsilon = 10$, the class imbalance and multicollinearity scenarios often fail to meet the median user's expectations.

Overall, this set of simulations meet's users' expectations according to these four metrics. Sign mismatch and significance mismatch are very sensitive to sample size and effect size. With 100,000 observations, even dramatic reductions in the effective sample size are not enough to drive sign mismatch and significance mismatch. This could easily change with fewer observations or SNR between 0 and 0.5. Large absolute relative errors when the SNR is 0 are somewhat artificial. When

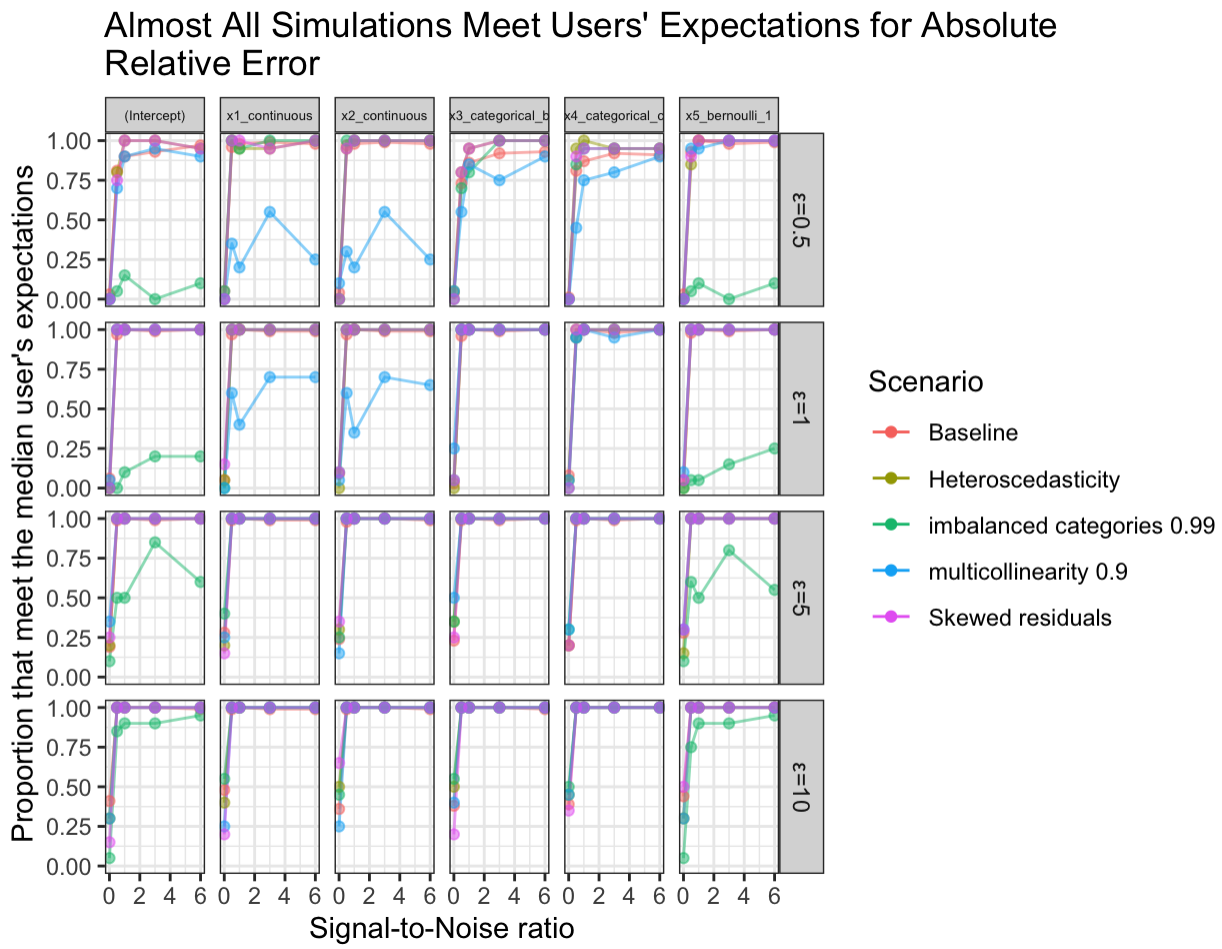

Figure 11: Each line represents the most extreme specification for a scenario. The results are limited to the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

the SNR is greater than 0, then absolute relative error is sensitive to effective sample size and these results would change with different sample sizes and scenarios. Confidence interval ratio is most sensitive to ϵ and SNR. The bootstrap confidence intervals have good coverage probabilities, so changes in confidence interval ratio are likely tied directly to the infusion of noise. The confidence interval ratio is directly tied to the idea of effective sample size.

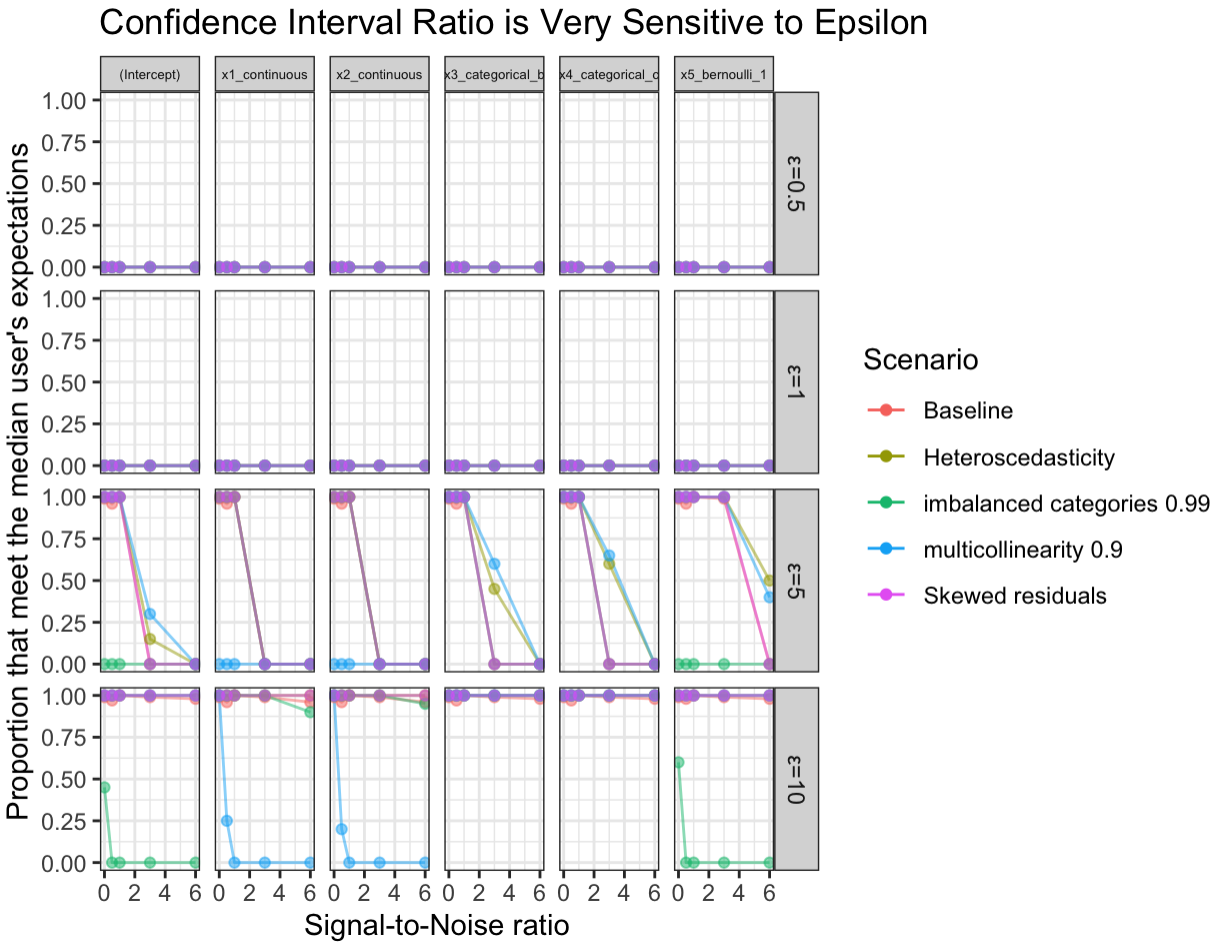

Figure 12: Each line represents the most extreme specification for a scenario. The results are limited to the Laplace mechanism, bootstrap confidence intervals, and DP range bounds.

4 Discussion

This paper introduces a framework for evaluating inferences from DP multiple linear regression methods on simulated data. This framework provides two key contributions. First, most privacy literature focuses the evaluation of DP regression methods on prediction instead of inference. Our framework provides useful tools for ensuring DP linear regression methods are useful to social scientists who focus on inference. Second, the framework could be useful for people who want to understand the error rates of a DP regression method before running the method. This would resemble power analysis but for statistical data privacy methods, and it would be particularly useful for users of a validation server who do not have access to the confidential data.

Our framework offers the novel contribution that we consider alternative scenarios where the data generating process does not perfectly satisfy the OLS assumptions. As two examples, we present a comparative analysis of results for alternative scenarios related to low-level frequencies and multicollinearity, using simulated datasets generated from a data-generating mechanism that maintains a consistent SNR. Different options exist for simulating many of the alternative scenarios we describe, and future work will be focused on exploring these different simulation strategies to assess their robustness and impact on findings.

We only explored a small subset of possible alternative scenarios. Future research will seek to extend this framework to consider other violations of normal linear model assumptions, such as non-linearity, independence of errors, and zero conditional mean assumption. Additionally, we will explore aspects like omitted variable bias, varying sample sizes and the number of predictors, incorporating different regularization strategies, and considering different privacy budgets to establish DP bounds for continuous unbounded variables. Moreover, we will introduce new utility metrics, such as the coverage of predictive intervals, to provide a more comprehensive evaluation framework.

Researchers interested in evaluating DP methods for regression via simulation studies must carefully consider how to set bounds for continuous unbounded variables. As observed in this simulation study, the different bounds significantly impacts method performance. Therefore, we recommend that simulation studies for regression analysis should involve unbounded continuous

variables and systematically test the robustness of results under various strategies for setting up bounds.

In general, the DP regression models we test do not perform well for inference unless the SNR is high and all the assumptions of OLS are satisfied. Under these conditions with 100,000 observations, the DP-based inference did not result in many mismatched signs or inferences. Applying these mechanisms to smaller data sets, which are common in the social sciences, would probably significantly impact inference under many privacy-loss budgets. It's striking that the quality of the results varied significantly with relatively modest changes to the scenarios. For example, the modest changes to the approach for generating sensitivities resulted in wildly different effective sample sizes. Outside of a simulation environment, it will be difficult for analysts and users of DP tools to anticipate the exact impact of DP noise on linear regression results, but the simulations help users get a much better idea than what theory might suggest.

Social scientists need methodologies like the ones evaluated in this paper that will allow them to evaluate "practical significance" while maintaining "practical privacy." In the future, we will test other proposed approaches, such as those relying on the observed/local sensitivity methods, such as Chetty and Friedman (2019), or DP Bayesian linear regression methods. Barrientos et al. (2024) initially did not incorporate local sensitivity based methods, because they focus on whether DP or formally private methods could produce accurate results. Barrientos et al. (2024) also did not evaluate DP Bayesian methods because they are still in early stages and are not as feasible to implement in a validation server framework. Hence, Barrientos et al. (2024) did not include Bayesian DP methods. However, as the literature provides promising new Bayesian methods that could be feasible, so we hope to expand this simulation study to include these new methods.

Finally, we adopt a user-centered approach where we evaluate the results against the expectations of potential users of results of DP regression. We believe any work putting DP into practice must tether evaluations to users' expectations. Follow-up work could involve collecting feedback from potential users about effective sample size, bias, and coverage probability, which we believe are more useful for evaluating results than sign mismatch, significance mismatch, and absolute relative error.

Acknowledgments

This research was funded by by the Alfred P. Sloan Foundation [G-2022-17149] and National Science Foundation National Center for Science and Engineering Statistics [49100422C0008].

We would like to thank our collaborators at SOI, especially Barry Johnson, Victoria Bryant, Chris Rexrode, Conrado Arroyo, Derek Gutierrez, and Giang Trinh for their amazing support.

We also thank our stellar validation server project team, consisting of Nikhita Airi, Leonard Burman, John Czajka, Surachai Khitatrakun, Graham MacDonald, Rob McClelland, Sybil Mendonca, Josh Miller, Livia Mucciolo, Madeline Pickens, Jean Clayton Seraphin, Deena Tamaroff, Silke Taylor, Erika Tyagi, and Doug Wissoker.

Thank you to Noah Johnson for reviewing our code and Madeline Pickens and Erika Tyagi for editing our paper.

Finally, we thank our advisory board for their advice and support. The members are John Abowd, Jim Cilke, Connie Citro, Jason DeBacker, Rick Evans, Dan Feenberg, Max Ghenis, Nick Hart, Matt Jensen, Ithai Lurie, Ashwin Machanavajjhala, Shelly Martinez, Robert Moffitt, Amy O'Hara, Mauricio Ortiz, Nancy Potok, Jerry Reiter, Rolando Rodriguez, Emmanuel Saez, Wade Shen, Aleksandra Slavković, Salil Vadhan, and Lars Vilhuber.

Author Contributions: We used the CRediT taxonomy[8] to indicate author contributions to this paper.

ARW: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, and Writing – review & editing

AFB: Conceptualization, Investigation, Methodology, Software, Writing – original draft, and Writing – review & editing

JS: Conceptualization, Methodology, Validation, and Writing – review & editing

CMB: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – original draft, and Writing – review & editing

[8]See website to learn more about the CRediT taxonomy, https://credit.niso.org

Conflict of Interest

The authors report there are no competing interests to declare.

References

Alabi, D., A. McMillan, J. Sarathy, A. Smith, and S. Vadhan (2022). Differentially private simple linear regression. *Proceedings on Privacy Enhancing Technologies 2*, 184–204.

Awan, J. and A. Slavković (2018). Differentially private uniformly most powerful tests for binomial data. *Advances in Neural Information Processing Systems 31*.

Balle, B. and Y.-X. Wang (2018). Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. pp. 394–403. Proceedings of Machine Learning Research: International Conference on Machine Learning.

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 (2024). A feasibility study of differentially private summary statistics and regression analyses with evaluations on administrative and survey data. *Journal of the American Statistical Association 119*(545), 52–65.

Bernstein, G. and D. R. Sheldon (2019). Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems 32*, 525–35.

Bowen, C. M., V. Bryant, L. Burman, J. Czajka, S. Khitatrakun, G. MacDonald, R. McClelland, L. Mucciolo, M. Pickens, K. Ueyama, et al. (2022). Synthetic individual income tax data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases*, pp. 191–204. Springer.

Bowen, C. M., V. Bryant, L. Burman, S. Khitatrakun, R. McClelland, P. Stallworth, K. Ueyama, and A. R. Williams (2020). A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases*, pp. 257–270. Springer.

Bowen, C. M., V. L. Bryant, L. Burman, S. Khitatrakun, R. McClelland, L. Mucciolo, M. Pickens, and A. R. Williams (2022). Synthetic individual income tax data: promises and challenges. *National Tax Journal 75*(4), 767–790.

Bryant, V. L., J. L. Czajka, G. Ivsin, and J. Nunns (2014). Design changes to the soi public use file (puf). In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, Volume 107, pp. 1–19. JSTOR.

Bun, M. and T. Steinke (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–58. Springer.

Chetty, R. and J. N. Friedman (2019). A practical method to reduce privacy loss when disclosing statistics based on small samples. In *AEA Papers and Proceedings*, Volume 109, pp. 414–420. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer.

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.

Dwork, C. and A. Roth (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*(3–4), 211–407.

Dwork, C. and G. N. Rothblum (2016). Concentrated differential privacy. *arXiv*.

Dwork, C., G. N. Rothblum, and S. Vadhan (2010). Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE.

Ferrando, C., S. Wang, and D. Sheldon (2020). General-purpose differentially-private confidence intervals. *Preprint, arXiv:2006.07749v1*.

Ferrando, C., S. Wang, and D. Sheldon (2021). General-purpose differentially-private confidence intervals. *arXiv preprint arXiv:2006.07749*.

Gillenwater, J., M. Joseph, and A. Kulesza (2021). Differentially private quantiles. *Preprint arXiv:2102.08244*.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, Volume 580. Springer.

Kasiviswanathan, S. P. and A. Smith (2014). On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality 6*(1).

Kifer, D. and A. Machanavajjhala (2011). No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 193–204.

Li, N., M. Lyu, D. Su, and W. Yang (2016). Differential privacy: From theory to practice. *Synthesis Lectures on Information Security, Privacy, & Trust 8*(4), 1–138.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008). Privacy: Theory meets practice on the map. *24th Institute of Electrical and Electronics Engineers International Conference on Intelligent Transportation Systems*, 277–286.

McSherry, F. D. (2009). Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 Association for Computing Machinery's Special Interest Group on Management of Data International Conference on Management of Data*, pp. 19–30. Association for Computing Machinery.

Mironov, I. (2017). Rényi differential privacy. In *Institute of Electrical and Electronics Engineers 30th Computer Security Foundations Symposium*, pp. 263–75. Institute of Electrical and Electronics Engineers.

Nagaraj, A. and M. Tranchero (2023). How does data access shape science? evidence from the impact of us census's research data centers on economics research. Technical report, National Bureau of Economic Research.

Nissim, K., S. Raskhodnikova, and A. Smith (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual Association for Computing Machinery Symposium on Theory of Computing*, pp. 75–84. Association for Computing Machinery.

Sheffet, O. (2017). Differentially private ordinary least squares. In *International Conference on Machine Learning*, pp. 3,105–114. Proceedings of Machine Learning Research.

Sheffet, O. (2019). Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pp. 789–827. Proceedings of Machine Learning Research.

Slavković, A. and J. Seeman (2023). Statistical data privacy: A song of privacy and utility. *Annual Review of Statistics and Its Application 10*, 189–218.

Snoke, J., C. M. Bowen, A. R. Williams, and A. F. Barrientos (2023). Incompatibilities between current practices in statistical data analysis and differential privacy. *arXiv preprint arXiv:2309.16703*.

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

Vu, D. and A. Slavkovic (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 138–143. IEEE.

Wang, Y., D. Kifer, and J. Lee (2019). Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality 9*(1).

Wang, Y.-X. (2018). Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. *Preprint, arXiv:1803.02596*.

Wasserman, L. and S. Zhou (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association 105*(489), 375–389.

Williams, A. R., J. Snoke, C. 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. *National Bureau of Economic Research*.

Appendix: Background on Differential Privacy

Differential privacy (DP) is a mathematical framework that provides a provable and quantifiable amount of privacy protection. There are several definitions and theorems of DP. In this section, we review two definitions of DP and three key theorems used in our simulation study. We use the following notation: $X \in \mathbb{R}^{n \times r}$ is the confidential dataset representing n data points and r variables and $M : \mathbb{R}^{n \times r} \to \mathbb{R}^k$ denotes the statistical query, i.e., M is a function mapping X to k real numbers. We denote randomized versions of M as $\mathcal{M}$ with the same domain and range. When appropriately parameterized, randomized mechanisms $\mathcal{M}$ can be said to implement particular DP frameworks.

Definitions of Differential Privacy

Definition 1. ***Differential Privacy*** *(Dwork et al., 2006): A sanitization algorithm, $\mathcal{M}$, satisfies ϵ-DP if for all subsets $S \subseteq Range(\mathcal{M})$ and for all X, X' such that $d(X, X') = 1$,*

$$\frac{\Pr(\mathcal{M}(X) \in S)}{\Pr(\mathcal{M}(X') \in S)} \leq \exp(\epsilon) \tag{5}$$

where $\epsilon > 0$ is the privacy loss budget and $d(X, X') = 1$ represents the possible ways that X' differs from X by one record.

Definition 1 is known as ϵ-DP. At a high level, DP links the potential for privacy loss to how much the answer of a query (e.g., statistic) is changed given the presence or absence of any possible person's data from any possible data set. The role of ϵ is to control the privacy loss. Intuitively, when ϵ decreases, the maximum distance between the probability distributions of $\mathcal{M}(X)$ and $\mathcal{M}(X')$ become smaller, indicating that $\mathcal{M}(X)$ and $\mathcal{M}(X')$ are less distinguishable in distribution. Hence, users cannot determine whether the mechanism's outputs are based on X or X', which in turn protects the confidential information of that record that distinguishes X and X'. Thus, low values of ϵ indicate high privacy levels and vice versa. ϵ-DP can also be interpreted from a more statistical perspective in the context of hypothesis testing and under both frequentist (Wasserman and Zhou, 2010) and Bayesian (Kasiviswanathan and Smith, 2014) paradigms.

Two definitions exist on what it means to differ by one record (Kifer and Machanavajjhala,

2011). One definition assumes the presence or absence of a record, where the dimensions of X and X' differ by one row, making X and X' unbounded neighbors. The other definition assumes a change in the value of one record, where X and X' have the same dimensions, making X and X' bounded neighbors. Kifer and Machanavajjhala (2011) refers to these as *unbounded DP* for presence or absence of a record and *bounded DP* for the change of a record. Li et al. (2016) state that unbounded DP satisfies an important composition theorem, which we will discuss later in this section (see Theorem 1), whereas bounded DP does not. In this paper, we assume unbounded DP, because we rely on Theorem 1.

Several relaxations of ϵ-DP have been developed in order to inject less noise into the output, such as (ϵ, δ)-DP (Dwork et al., 2006), probabilistic DP (Machanavajjhala et al., 2008), concentrated DP (Dwork and Rothblum, 2016), Rényi DP (Mironov, 2017), and zero-concentrated DP (Bun and Steinke, 2016). Although these definitions use the same provable privacy framework, they utilize alternative parameters offering different privacy guarantees. In return, they allow more possibilities for the type of noise added. We only review probabilistic DP, because it is the only other formally private definition used in our simulation study.

Definition 2. (ϵ, δ)***-Differential Privacy*** *(Dwork et al., 2006): A sanitization algorithm, $\mathcal{M}$, satisfies (ϵ, δ)-DP if for all X, X' that are $d(X, X') = 1$,*

$$\Pr(\mathcal{M}(X) \in S) \leq \exp(\epsilon) \Pr(\mathcal{M}(X') \in S) + \delta \tag{6}$$

where $\delta \in [0, 1]$.

Definition 2 provides a simple relaxation of Definition 1 by adding the parameter δ, which allows the privacy loss associated with the ϵ bound to fail at a rate no greater than δ. Definition 1 can also be defined as a special case of (ϵ, δ)-DP when $\delta = 0$.

Global Sensitivity and DP Mechanisms

In this section, we introduce the concept of global sensitivity and present three of the fundamental mechanisms that satisfy ϵ-DP and (ϵ, δ)-DP and form the building blocks of the DP algorithms we test in this paper.

Independent of the values of ϵ and δ, an algorithm that satisfies ϵ-DP or (ϵ, δ)-DP must adjust the amount of noise added to the output based on the maximum possible change between any two databases that differ by one row. This is commonly referred to as the global sensitivity (GS), given in Definition 3.

Definition 3. *l_1-Global Sensitivity (Dwork et al., 2006): For all X, X' such that $d(X, X') = 1$, the global sensitivity of a function M is*

$$\Delta_1(M) = \sup_{d(X,X')=1} \|M(X) - M(X')\|_1 \tag{7}$$

We can calculate global sensitivity under different norms. For instance, $\Delta_2(M)$ represents the l_2 norm GS (l_2-GS) of the function M. Although the definition is straightforward, calculating the GS can often be difficult in practice. For instance, we cannot calculate a finite GS of one of the most common statistics, the sample mean, if the variable is not bounded (or the bound is not known).

A commonly used mechanism satisfying ϵ-DP is the Laplace mechanism, given in Definition 4. Dwork et al. (2006) proved it satisfies ϵ-DP and uses the l_1-GS. Another popular mechanism is the Gaussian mechanism, given in Definition 5, which uses the l_2-GS of the statistical query. Dwork and Roth (2014) showed the Gaussian mechanism satisfies (ϵ, δ)-DP.

Definition 4. *Laplace mechanism (Dwork et al., 2006): Given any function $M : \mathbb{R}^{n\times r} \to \mathbb{R}^k$, the Laplace mechanism is defined as:*

$$\mathcal{M}(X) = M(X) + (\eta_1, \ldots, \eta_k). \tag{8}$$

where $(\eta_1, \ldots, \eta_k)$ are i.i.d. Laplace$(0, \frac{\Delta_1(M)}{\epsilon})$.

Definition 5. *Gaussian mechanism (Dwork and Roth, 2014): Given any function $M : \mathbb{R}^{n\times r} \to \mathbb{R}^k$, the Gaussian mechanism is defined as:*

$$\mathcal{M}(X) = M(X) + (\eta_1, \ldots, \eta_k). \tag{9}$$

where $(\eta_1, \ldots, \eta_k)$ *are i.i.d.* $N\left(0, \sigma^2 = \left(\frac{\Delta_2(M)\sqrt{2\log(1.25/\delta)}}{\epsilon}\right)^2\right)$.

Composition and Post-processing Theorems

Lastly, we introduce the important concepts of composition and post-processing. These enable the development of more complex algorithms that combine DP mechanisms with post-processing to release multiple statistics with additional structural or noise-reducing enhancements. The composition theorems given in Theorem 1 formalize the concept of totaling the privacy loss incurred across multiple queries or datasets.

Theorem 1. ***Composition Theorems*** *(Bun and Steinke, 2016; Dwork and Rothblum, 2016; McSherry, 2009): Suppose a mechanism,* $\mathcal{M}_j$*, provides* (ϵ_j, δ_j)*-DP for* $j = 1, \ldots, J$*.*

a) ***Sequential Composition:*** *The sequence of* $\mathcal{M}_j(X)$ *applied on the same* X *provides* $(\sum_{j=1}^{J} \epsilon_j, \sum_{j=1}^{J} \delta_j)$*-DP.*

b) ***Parallel Composition:*** *Let* X_j *be disjoint subsets of the dataset* X*,* $j = 1, \ldots, J$*. The sequence of* $\mathcal{M}_j(X_j)$ *provides* $(\max_{j\in\{1,\ldots,J\}} \epsilon_j, \max_{j\in\{1,\ldots,J\}} \delta_j)$*-DP.*

If we want to make J statistical queries on X and we want the total privacy loss to equal ϵ, the composition theorems state under what conditions we may allocate portions of the overall ϵ to each statistic. Under sequential composition, a typical appropriation is dividing ϵ and δ equally by J. For example, a data practitioner might want to query the mean and standard deviation of a variable. These two queries will require using the sequential composition, allocating an equal amount of privacy budget to each query. Dwork et al. (2010) proposed other forms of sequential composition, but the methods we test do not rely on these works.

Conversely, parallel composition does not require splitting the budget because the noise is applied to disjoint subsets of the input domain. For example, privacy experts will often leverage parallel composition to sanitize histogram counts, assuming that the bins are disjoint subsets of the data. Noise can then be added to each bin independently without needing to split ϵ or δ.

Theorem 2 (post-processing) states that any function applied to the output of a DP mechanism also satisfies DP. Some DP methods, as will be shown later, use the post-processing theorem to

correct any inconsistencies or values that are not possible and to compute additional summaries required to perform statistical inference.

Theorem 2. ***Post-Processing Theorem*** *(Bun and Steinke, 2016; Dwork et al., 2006; Nissim et al., 2007): If $\mathcal{M}$ be a mechanism that satisfies (ϵ, δ)-DP and g be any function, then $g(\mathcal{M}(X))$ also satisfies (ϵ, δ)-DP.*