

A Feasibility Study of Differentially Private Summary Statistics and Regression Analyses for Administrative Tax Data

Andrés F. Barrientos¹, Aaron R. Williams², Joshua Snoke³, and Claire McKay Bowen²

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

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

³*RAND Corporation, jsnoke@rand.org*

Keywords— differential privacy, validation server, administrative data, tax analysis

Abstract: Federal administrative tax data are invaluable for research, but because of privacy concerns, access to these data is typically limited to select agencies and a few individuals. An alternative to sharing microlevel data are validation servers, which allow individuals to query statistics without accessing the confidential data. This paper studies the feasibility of using differentially private (DP) methods to implement such a server. We provide an extensive study on existing DP methods for releasing tabular statistics, means, quantiles, and regression estimates. We also include new methodological adaptations to existing DP regression algorithms for using new data types and returning standard error estimates. We evaluate the selected methods based on the accuracy of the output for statistical analyses, using real administrative tax data obtained from the Internal Revenue Service Statistics of Income (SOI) Division. Our findings show that a validation server would be feasible for simple statistics but would struggle to produce accurate regression estimates and confidence intervals. We outline challenges and offer recommendations for future work on validation servers. This is the first comprehensive statistical study of DP methodology on a real, complex dataset, that has significant implications for the direction of a growing research field.

1 Introduction

Federal tax data, derived from individuals’ and businesses’ tax and information returns, are invaluable resources for research on a range of topics. Such research improves our understanding of individuals’ and firms’ responses to economic incentives, and researchers can also use the data to

study areas far removed from taxation. For example, [Chetty et al. \(2014\)](#) used tax data to study economic mobility across generations and how elementary school teacher quality affects economic outcomes later in life. However, full access to these data is available only to select government agencies, to a very limited number of researchers working in collaboration with analysts in those agencies, or through highly selective programs within the Internal Revenue Service (IRS) Statistics of Income (SOI) Division. In addition, the existing process of manually vetting each statistical release for disclosure risks is labor intensive and imperfect because it relies on subjective human review. The tremendous demand to participate in such projects, which is limited by SOI resource constraints, indicates that more high-quality research could be conducted if a safe and less resource-intensive method were developed to expand access.

1.1 Background on Accessing Confidential Data

At the IRS, the current process to release analytic results on confidential datasets requires researchers to undergo an extensive background check to access the data. IRS staff then must review any results the researchers want to release. This process reflects the norm for researchers wishing to access federal, confidential data. In general, researchers either gain access from a public use file that is an altered version of the confidential data or have direct access to the confidential data.

As a potential middle ground between the two extremes, the U.S. Census Bureau provides research access to two experimental synthetic databases via the Synthetic Data Server at Cornell University: the Synthetic Longitudinal Business Database and the Survey of Income and Program Participation’s Synthetic Beta Data Product ([Benedetto et al., 2013](#); [Drechsler and Vilhuber, 2014](#)). The Synthetic Data Server hosts a validation server that allows researchers to submit their statistical programs to run on the underlying administrative data after testing it on the publicly available synthetic data. However, this server has two disadvantages. First, because it is not automated, the process consumes limited staff time, which demand often exceeds. This situation causes long delays for approval. Second, reviews may be inconsistent because they are manually evaluated by humans and do not adhere to formal notions of privacy that constrain the allowable output.

To address these problems, some privacy researchers proposed a newer privacy loss definition, differential privacy (DP), to speed up the process and to remove the ad-hoc decisions ([Dwork](#)

et al., 2006). Over the last decade, many data privacy experts have come to regard DP as the gold standard for privacy protection. It is often called a *formally private* method because people can mathematically prove the privacy loss from a data publication that uses DP methods. These methods differ from prior statistical disclosure control or limitation methods because they do not require the same strong assumptions on how much information an intruder may have or what kind of disclosure is likely to occur. Note that this does not imply DP protects from all attacks, rather, for a defined type of privacy loss it offers provable amounts of protection.

At a high level, DP links the potential for privacy loss to how much the answer of a query (such as a statistic) is changed given the absence or presence of the most extreme possible person or observation in the data population. DP requires the level of protection is set proportionally to this maximum potential change, thereby providing formal privacy protections scaled to the worst-case scenario. For further details, [Dwork and Roth \(2014\)](#) provide a rigorous mathematical review of DP. [Bowen and Garfinkel \(2021\)](#) cover the basics of DP and its challenges for adoption geared toward a general, mathematical audience, whereas [Nissim et al. \(2017\)](#) and [Snoke and Bowen \(2019\)](#) describe DP for a nontechnical, general audience.

A validation server that releases a DP output could theoretically be automated because every released statistic would meet a pre-defined privacy guarantee. This could reduce the burden of manual review that exists in current validation servers and enable increased access to the server, allowing researcher on a wider range of topics that use IRS tax data. In a similar but different approach, [Barrientos et al. \(2018\)](#) created a pilot *verification* server for the U.S. Office of Personnel Management that allowed users to verify synthetic data estimates while satisfying DP.

1.2 Contributions from this Paper

To understand the feasibility of creating an automated validation server, we conduct an extensive feasibility study on state-of-the-art DP methods for releasing tabular statistics, mean and quantile statistics, and regression analyses with cross-sectional data. Based on the current DP literature and informal interviews with tax experts, we prioritized these analyses as a first stage of a potential validation server. There are several other analyses, such as model selection or regression discontinuity design, that have been identified as important but the current DP methodology is at an early

stage and would not support a validation server implementation. We will explore these in future work for the development of a validation server.

To measure feasibility, we test the DP methods on the SOI Public Use File (PUF). SOI annually develops and releases this database of sampled individual income tax returns with privacy protections. Several organizations, such as the American Enterprise Institute, the Urban-Brookings Tax Policy Center, and the National Bureau of Economic Research develop PUF-based microsimulation models that help inform the public on potential impacts of policy proposals. But access to this public file is limited to certain institutions, and we cannot provide the full data for others to replicate our results in this paper. For this reason, we also test on the 1994 to 1996 Current Population Survey Annual Social and Economic Supplements (CPS ASEC), publicly accessible through IPUMS USA [Ruggles et al. \(2021\)](#). Crucially, it has similar variables as the PUF, and the case study results on the CPS ASEC will be similar to those on the SOI. We provide results from the SOI in this paper and results from the CPS ASEC in the Supplemental Materials, along with a public repository of the code, data, and results. Testing these methods on specific case studies is a significant contribution because it allows assessing how robust and accurate these methods are under conditions commonly encountered in real-life applications. Proposed methods (particularly for regression) have mostly been tested without paying attention to specific applied contexts. The assessment of their performance usually focuses on verifying that their outputs become more accurate, without considering a feasibility threshold for accuracy, as the sample size increases and less privacy protection is required.

When selecting which DP algorithms to use, we provide descriptions of each algorithm, consider their ease of implementation, determine whether they require any additional tuning parameters, assess their computational feasibility, and other advantages and disadvantages. As part of study, we note which methods work in theory under specific conditions that would not normally be met in practice. This assessment will be useful for the data privacy and confidentiality community in developing better and more robust DP algorithms in the future.

We also contribute new methodology for some DP regression methods. We improve the sensitivity calculation for fitting the models with binary or categorical predictors and provide standard

error estimates in addition to point estimates. Many of the existing methods for DP regression only provide point estimates, but we require that the validation server provides full inference. Further, obtaining standard errors or confidence intervals is crucial to most statistical analyses.

We define feasibility based on the impact of DP mechanisms on analyses for making public policy decisions and their accuracy according to several utility metrics. We evaluate the methods using real data and identify how specific data features might challenge some of these methods. To the best of our knowledge, this paper covers the first comprehensive case study on DP methodology for various statistical analyses. DP is a rapidly growing and popular field of study, but the vast majority of the work has focused solely on theoretical developments. Our results and conclusions will be vital in informing the current state of DP methods, addressing practical problems, and identifying directions for future work, particularly for statistical inference.

We organize the remainder of the paper as follows. Section 2 reviews the definitions, concepts, and theorems of differential privacy and the fundamental differentially private mechanisms that we use. Section 3 examines several differentially private methods, providing a thorough discussion on which methods we tested, the extensions we provided to the methods, and which methods could not be implemented in practice. Section 4 compares the differentially private methods on IRS tax data and determines the feasibility of using these algorithms for a validation server. Concluding remarks and areas for future work are given in Section 6. Additional technical details on the algorithms and expanded results for our case studies can be found in the Supplementary Materials.

2 Differential Privacy

Differential privacy (DP) offers a provable and quantifiable amount of privacy protection, colloquially referred to as the privacy loss budget. Those in the data privacy and confidentiality community should note that DP provides a statement about the algorithm (or mechanism), not the data—a common misconception. In other words, DP requires that the *mechanism* or *algorithm* provably meets the privacy definition. We refer to these methods as differentially private (DP) mechanisms or algorithms. Note that we use DP to mean both differential privacy and differentially private.

For this section, we reproduce the pertinent definitions and theorems of DP with the following

notation: $X \in \mathbb{R}^{n \times r}$ is the original dataset representing n data points and r variables and $M : \mathbb{R}^{n \times r} \rightarrow \mathbb{R}^k$ denotes the statistical query, i.e., M is a function mapping X to k real numbers.

2.1 Definitions and Theorems

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

$$\frac{\Pr(\mathcal{M}(X) \in S)}{\Pr(\mathcal{M}(X') \in S)} \leq \exp(\epsilon) \quad (1)$$

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 provides what is known as ϵ -DP. There are varying understandings of what it means to differ by one record. One interpretation is the presence or absence of a record, and the other has the difference as a change, where X and X' have the same dimensions. Li et al. (2016) refers to these interpretations as *unbounded DP* for addition or removal of a record and *bounded DP* for the change of a record. They prove that unbounded DP satisfies an important composition theorem we will discuss later in this section (see Theorem 1), whereas bounded DP does not. Because many DP methods rely on Theorem 1, we assume unbounded DP in this paper.

Several relaxations of ϵ -DP have been developed in order to inject less noise into the output, such as approximate DP (Dwork et al., 2006), probabilistic DP (Machanavajjhala et al., 2008), concentrated DP (Dwork and Rothblum, 2016), Rényi differential privacy (Mironov, 2017), and zero-concentrated DP (Bun and Steinke, 2016). Though these definitions are still formally private, they offer slightly weaker privacy guarantees. In return, they typically lessen the amount of noise required. We will cover approximate DP, also known as (ϵ, δ) -DP, and zero-concentrated DP in depth, because most of the methods we test in our study use one of these two definitions.

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 $d(X, X') = 1$,

$$\Pr(\mathcal{M}(X) \in S) \leq \exp(\epsilon) \Pr(\mathcal{M}(X') \in S) + \delta \quad (2)$$

where $\delta \in [0, 1]$. ϵ -DP is a special case of (ϵ, δ) -DP when $\delta = 0$.

Definition 2 provides a simple relaxation of Definition 1 by adding the parameter δ . This allows, with small probability, that the strict bound given does not hold, which can be useful when dealing with extreme yet very unlikely cases.

Dwork and Rothblum (2016) proposed concentrated DP, which aimed to reduce the privacy loss over multiple computations (more on composition of multiple queries soon when discussing Theorem 1). This definition of privacy was later on improved by Bun and Steinke (2016) who introduced zero-concentrated DP (zCDP or ρ -zCDP), given in Definition 3. They also show in their Proposition 1.3 that if M satisfies ρ -zCDP, then M is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$ -DP for any $\delta > 0$. For the other direction, their Proposition 1.4 states that if M satisfies ϵ -DP, then M satisfies $(1/2\epsilon^2)$ -zCDP, which allows us to relate ρ -zCDP algorithms to an ϵ -DP equivalent.

Definition 3. Zero-Concentrated Differential Privacy (Bun and Steinke, 2016): A sanitization algorithm, \mathcal{M} , satisfies (ξ, ρ) -zero-concentrated differential privacy if for all X, X' that are $d(X, X') = 1$ and $\alpha \in (1, \infty)$,

$$D_\alpha(\mathcal{M}(X) || \mathcal{M}(X')) \leq \xi + \rho\alpha, \quad (3)$$

where $D_\alpha(\mathcal{M}(X) || \mathcal{M}(X'))$ is the α -Rényi divergence between the distribution of $\mathcal{M}(X)$ and the distribution of $\mathcal{M}(X')$, ξ and ρ are positive constants, and $\alpha \in (1, \infty)$.

As mentioned before, many DP algorithms require repeated responses from a query system, such as a validation server. Each time a statistic or output is released, data information “leaks” and must be protected. DP protects the information by splitting the amount of ϵ used for each output, and the composition theorems formalize this concept.

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 or (ξ_j, ρ_j) -zCDP for $j = 1, \dots, J$.

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 or $(\sum_{j=1}^J \xi_j, \sum_{j=1}^J \rho_j)$ -zCDP.

b) **Parallel Composition:** Let D_j be disjoint subsets of the input domain D . The sequence of $\mathcal{M}_j(X \cap D_j)$ provides $(\max_{j \in \{1, \dots, J\}} \epsilon_j, \max_{j \in \{1, \dots, J\}} \delta_j)$ -DP or $(\max_{j \in \{1, \dots, J\}} \xi_j, \max_{j \in \{1, \dots, J\}} \rho_j)$ -zCDP.

More simply, suppose there are J many statistical queries on X . The composition theorems state that we may allocate a portion of the overall desired level of ϵ to each statistic by sequential composition. A typical appropriation is dividing ϵ 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. Conversely, parallel composition does not require splitting the budget because the noise is applied to disjoint subsets of the input domain. Privacy experts will often leverage parallel composition, for instance, to sanitize histogram counts, where the bins are disjoint subsets of the data. In this example, noise can be added to each bin independently without needing to split ϵ .

The post-processing theorem is another important theorem, which states that any function applied to a DP output also satisfies DP. Many DP methods 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 or (ξ, ρ) -zCDP, and g be any function, then $g(\mathcal{M}(X))$ also satisfies (ϵ, δ) -DP or (ξ, ρ) -zCDP.

2.2 Differentially Private Mechanisms

In this section, we present two of the fundamental mechanisms we consider that satisfy ϵ -DP and (ϵ, δ) -DP. We employ additional mechanisms that will be discussed in Section 3. For a given value of ϵ and δ , an algorithm that satisfies DP or approximate DP will adjust the amount of noise added to the output based on the maximum possible change between two databases that differ by one row. This value is commonly referred to as the global sensitivity (GS), given in Definition 4.

Definition 4. 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

$$\Delta_1(M) = \sup_{d(X, X')=1} \|M(X) - M(X')\|_1 \quad (4)$$

We can calculate sensitivity under different norms. For instance, $\Delta_2(M)$ represents the l_2 norm GS, l_2 -GS, of the function M . Another way of thinking about the GS is that it measures the statistical query's robustness to outliers. Though the definition is straightforward, calculating the GS can often be difficult. For instance, we cannot directly calculate an upper bound for the GS of one of the most common statistical analyses, regression, where the coefficients are unbounded. To address this issue, privacy researchers had to be creative in figuring out how to add noise to regression analyses. We discuss this further in Section 3.4.

The most basic mechanism satisfying ϵ -DP is the Laplace mechanism, given in Definition 5, and was first introduced by Dwork et al. (2006). Another popular mechanism is the Gaussian mechanism that satisfies (ϵ, δ) -DP, given in Definition 6, which uses the l_2 -GS of the statistical query.

Definition 5. Laplace mechanism (Dwork et al., 2006): The Laplace mechanism satisfies ϵ -DP by adding noise (η_1, \dots, η_k) to M that are independently drawn from a Laplace distribution with the location parameter at 0 and scale parameter of $\Delta_1(M)\epsilon^{-1}$ such that

$$\mathcal{M}(X) = M(X) + (\eta_1, \dots, \eta_k). \quad (5)$$

Definition 6. Gaussian mechanism (Dwork and Roth, 2014): The Gaussian mechanism satisfies (ϵ, δ) -DP by adding Gaussian noise with zero mean and variance, σ^2 , such that

$$\mathcal{M}(X) = M(X) + (\eta_1, \dots, \eta_k) \quad (6)$$

where η_1, \dots, η_k are independently drawn and $\sigma = \Delta_2(M)\epsilon^{-1}\sqrt{2\log(1.25/\delta)}$.

Although Dwork and Roth (2014) proposed Gaussian mechanism for (ϵ, δ) -DP, the Gaussian

mechanism satisfies $((\Delta_2(M))^2/2\sigma^2)$ -zCDP, per Proposition 1.6 from [Bun and Steinke \(2016\)](#). Additionally, if the l_2 -GS is 1 (which is true for all counting queries), then the Gaussian mechanism satisfies $(\alpha, \frac{\alpha}{2\sigma^2})$ -Rényi DP, per Corollary 3 from [Mironov \(2017\)](#). We use this relationship for multiple counting queries to reduce the amount of noise being added from the Gaussian mechanism.

Both the Laplace and Gaussian mechanisms are simple and quick to implement, but they apply only to numerical values (without additional post-processing, Theorem 2). A more general ϵ -DP mechanism is the Exponential mechanism, given in Definition 7, which allows for the sampling of values from a noisy distribution rather than adding noise directly.

Definition 7. *Exponential mechanism* ([McSherry and Talwar, 2007](#)): *The Exponential mechanism releases values with a probability proportional to*

$$\exp\left(\frac{\epsilon u(X, \theta)}{2\Delta_1(u)}\right) \quad (7)$$

and satisfies ϵ -DP, where $u(X, \theta)$ is the score or quality function that determines the values for each possible output, θ , on X .

2.3 Setting the Privacy Loss Budget

The scientific community still has no general consensus on what value of ϵ should be used for practical implementation. Early differential privacy research focused on ϵ values that were less than or equal to one and suggested that an epsilon of two or three would release too much information ([Dwork, 2008](#)). Other technical interpretations relating to hypothesis testing ([Wasserman and Zhou, 2010](#)) or odds-ratios ([Machanavajjhala et al., 2008](#)) have also been proposed as ways of interpreting and setting limits on ϵ .

More recently, many privacy researchers working in practical applications frame the decision as a social and policy question. They interpret the parameter as a way to quantify the trade-off between accuracy and worst-case privacy loss. Privacy experts should inform and explain to policymakers potential ways to interpret the privacy and utility trade-off. But, ultimately, stakeholders will need to set the privacy parameter in ways that are relevant to their contexts.

Accordingly, many practical applications of differential privacy use large values of ϵ (by the-

oretical standards). In 2008, for example, privacy researchers applied (ϵ, δ) -differential privacy method with values at $(8.6, 10^{-5})$ to release a synthetic version of the OnTheMap data, a United States commuter dataset (Machanavajjhala et al., 2008). More recently, in 2020, Google’s COVID-19 Mobility Reports used 2.64-differential privacy for the daily reports a total of 79.22-differential privacy monthly (Aktay et al., 2020). In the same year, LinkedIn revealed their LinkedIn’s Audience Engagement API that protected LinkedIn members’ content engagement data, which used (ϵ, δ) -differential privacy with daily values of $(0.15, 10^{-10})$ or $(34.9, 7 \times 10^{-9})$ for monthly queries (Rogers et al., 2020). In 2021, the Census Bureau committed to ϵ values of 17.14 for the persons file and 2.47 for the housing units data on the 2020 Census.¹

Given these examples and the evolving understanding of ϵ , we explore a wide range of ϵ values in this paper based on values seen in theoretical work and practical applications. By doing so, we will gain a better sense of the feasibility of methods at different levels of ϵ . Specifically, we will examine the effects of ϵ and δ on accuracy within the context of our study, which we hope will contribute to the conversation about how to set the privacy-loss budget.

It is also important to understand that when considering the total privacy-loss budget, choosing the value of ϵ for any single query is sensitive to other factors, such as the sample size, the total number of desired queries, and the size of the population from which data are sampled. Those questions concern the overall framework that would need to be created to enable a fully implemented validation server, and we do not seek to answer those questions in this paper. We make this point to note that our goal is simply to identify which mechanisms are desirable, if any, based on their relative privacy and utility trade-offs, rather than their absolute trade-offs.

3 Differentially Private Algorithms for a Tax Data Validation Server

3.1 Tabular Statistics

The literature has shown that adding Laplace noise produces the most accurate estimates for single tabular queries. For example, (Bowen et al., 2021; Liu, 2018; Rinott et al., 2018; Shlomo, 2018) found the Laplace mechanism is still “hard to beat” for disseminating frequency tables, when

¹The U.S. Census Bureau set the final privacy loss budget on June 9, 2021, which can be found at <https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html>.

the data have a large number of observations, or there are a lot of parameters and attributes to consider. We also consider the Gaussian mechanism, since adding Gaussian noise may perform better than adding Laplace noise for multiple counting queries because of how the tails compose (Wang et al., 2019). We can take advantage of this composition property by leveraging Rényi differential privacy (Mironov, 2017), which reduces the amount of noise added to each count. A drawback to using the Gaussian mechanism is that it satisfies (ϵ, δ) -DP, requiring the data user to balance two privacy parameters instead of one. There are other DP algorithms that tackle more complex tabular statistics, but for this study we are only concerned with simple counting queries.

3.2 Quantile Statistics

The leading methods for generating differentially private quantiles use either the Laplace mechanism or the Exponential mechanism. For instance, Smith (2011) proposed an algorithm, *IndExp*, for selecting individual quantiles using the Exponential mechanism. *IndExp* has since been implemented in the SmartNoise (2020) and IBM (2019) DP libraries. This method was recently extended by Gillenwater et al. (2021) to two other algorithms, *AppIndExp* and *JointExp*. The former is the same as *IndExp* but uses the composition theorem from Dong et al. (2020) to choose an optimal ϵ for multiple queries given a total (ϵ, δ) , and the latter samples multiple quantiles jointly. Using the Laplace mechanism, Nissim et al. (2007) developed an approach for sampling median values using smooth sensitivity that can be extended to query any other quantile. We will hereafter refer to this method as *Smooth*. A few other approaches or adaptations of the approaches listed here have been proposed, but we did not test them because they require fine-tuning based on distribution assumptions that a researcher might not realistically have in our validation server setting.

3.3 Means and Confidence Intervals

For mean estimates, we reviewed DP methods that released means with their associated confidence intervals (CI). When we explored the general literature, we found common approaches use the Laplace mechanism, Gaussian mechanism, or Exponential mechanism for releasing some means with CIs. Du et al. (2020) conducted a comprehensive research study that aimed to move the theory to practice for releasing DP CIs. The authors developed five new methods, two based on directly applying Laplace noise named *NOISYVAR* and *NOISYMAD* and three based on querying quantiles

from the Exponential mechanism to estimate the standard deviation named *CENQ*, *SYMQ*, and *MOD*. [Du et al. \(2020\)](#) also compared their methods against [Karwa and Vadhan \(2017\)](#); [D’Orazio et al. \(2015\)](#); and [Brawner and Honaker \(2018\)](#).

A potential limitation to the approaches using quantiles is that they strongly assumes the data are approximately Gaussian. We chose to not test these methods on heavily skewed data because preliminary tests showed poor performance. Also, some methods require more information than is realistic for our application. For instance, [Biswas et al. \(2020\)](#); [Bowen and Liu \(2020\)](#); [D’Orazio et al. \(2015\)](#); [Karwa and Vadhan \(2017\)](#) require the researcher to set bounds on the standard deviation to calculate the GS. In a validation server setting, a researcher might not have a good sense of the bounds. Given this limitation, we did not test them for our study.

3.4 Differentially Private Regression Analyses

We begin with an overview of the currently available DP approaches for regression analyses. We then explain the criteria for including methods in this feasibility study, discuss the selected methods, and detail any adaptations required to include the methods in the experiments.

3.4.1 Traditional Differentially Private Approaches for Regression Analyses

We classify DP methods for regression analysis according to the outputs they produce: (1) point estimates only, (2) point and interval estimates, and (3) other outputs related to regression analysis, such as diagnostic plots. Because we are particularly interested in methods that provide full statistical inference, we focus our study on methods from category (2) and discuss these methods in greater detail. We survey the other two types further in the Supplemental Materials.

[Sheffet \(2017\)](#) developed (ϵ, δ) -differential private algorithms that, with certain probability, output summary statistics useful for either traditional, linear regression or ridge regression. When the outputs are summaries for ridge regression, the penalization parameter is a function of the algorithm’s inputs instead of being predefined by the user. [Sheffet \(2017\)](#) derives CIs and t -statistics that account for the noise added to the confidential summaries and shows how such statistics relate to the underlying truth. In addition, the author discusses a different algorithm that adds Gaussian noise directly to the sufficient statistics and shows that users could obtain

CIs for the regression coefficients under certain conditions (i.e., the norm of the true regression coefficients is upper bounded). Despite this method’s promise, the paper does not provide practical guidance on defining additional tuning parameters that are essential to guaranteeing the correct confidence or significance level of the CIs and t-statistics. The lack of publicly available code also contributes to excluding it from our study.

[Sheffet \(2019\)](#) proposed a follow-up (ϵ, δ) -DP mechanism that provides estimates with random noise drawn from the Wishart distribution. The mechanism defines a noisy statistic that preserves the property of being positive-definite; a common problem when adding noise to the regression sufficient statistics. But, this mechanism only works for $\epsilon < 1$. [Wang et al. \(2019\)](#) also developed (ϵ, δ) - and ϵ -DP methods that release noisy versions of the summary statistics while preserving positive definiteness. These methods add noise using either a normal distribution (for (ϵ, δ) -differentially privacy) or the spherical analogue of the Laplace distribution (for ϵ -differential privacy). The positive definiteness is achieved by using eigenvalue decomposition and censoring the eigenvalues falling below a given threshold. Although [Sheffet \(2019\)](#) and [Wang et al. \(2019\)](#) do not derive CIs or t-values under the proposed mechanisms for the normal linear model, these contributions paved the way to develop DP methods that allow full inference for regression coefficients.

[Ferrando et al. \(2020\)](#) developed a general approach to produce point and interval estimates for different DP mechanisms, including linear regression. The paper outlines two ϵ -DP strategies that employ a noisy version of the sufficient statistics. The first one applies the noisy statistics in classic ordinary least squares point and interval estimators. This strategy’s accuracy and coverage is ensured by large-sample arguments. The second strategy also uses plug-in estimators, but computes CIs by means of parametric bootstrap. The method accounts for the injected noise as well as the underlying sampling distribution. A drawback of these two approaches is the lack of clarity on computing the points and interval estimates of the coefficients when the inverse of the noisy covariance matrix is not positive-definite. We employ this method in the feasibility study and apply a regularized version of the noisy sufficient statistic to handle the positive definiteness issue. Details on how we use regularization are provided in [Section 3.4.2](#). We acknowledge that a new version of [Ferrando et al. \(2020\)](#) has been recently released, where the authors made the update after

we completed our feasibility study (Ferrando et al., 2021). Although there are slight differences between the main algorithms in each version, the first version algorithm produces valid results. We provide a full description of the algorithm we used in the Supplementary Materials.

Bernstein and Sheldon (2019) offered an approach that relies on a noisy version of the sufficient statistics. The procedure uses a Bayesian framework and employs a large-sample distributional characterization of the sufficient statistics. Using a Bayesian approach allows the authors to draw from the regression coefficient’s posterior distribution and, thus, provide point and interval estimates. Wang (2018) also provided a method that draws from the posterior distribution of the regression coefficients. However, users need to spend part of the privacy budget for each draw. This aspect limits the applicability of Wang (2018)’s algorithm, since any accurate Monte Carlo approximations would divide the privacy budget into many small values. Splitting the privacy parameter too much dramatically decreases the method’s statistical usefulness. Tax economists rarely use Bayesian methods, so we don’t expect many users will be familiar with these models. Because of this, we didn’t pursue DP Bayesian regression methods further for this case study. Nonetheless, we plan to consider Bayesian approaches in future versions of the validation server.

3.4.2 Selected Methods and Adaptations

When selecting methods to include in the study, we sought to maximize the statistical usefulness of the outputs and the feasibility of implementation. We select methods that are frequentist, can be used for linear regression models with normal errors, can handle multiple predictors, and provide a full inference. For example, we exclude methods that only provide t-values for regression coefficients without also providing point estimates. Finally, we exclude procedures that meet the criteria above yet were not possible to implement, notably if: the manuscript lacks the information needed to implement the method it describes, the pseudocode is absent and implementation requires non-trivial choices reflecting the theory underlying the proposed method, the method has difficult-to-fix errors, the authors failed to reply to inquiries about their method or its implementation, or the method achieves differential privacy under nontestable assumptions.

We assess each of the methods in Section 3.4.1 based on the selection criteria listed in the previous subsection. Without our additional adaptations, only one method (that of Ferrando et al.

(2020)) met all the inclusion criteria to be implemented in a validation server. By making adaptations to the methods in Ferrando et al. (2020) and applying them to other methods, we increased the number of testable methods from 1 to 6.

We included Brawner and Honaker (2018)’s method, because, even though it was not originally designed for linear regression, a small adaptation would make it eligible. We modified both Ferrando et al. (2020)’s and Brawner and Honaker (2018)’s methods and obtained new versions of the methods to compare in the feasibility study. Importantly, we repurposed elements of the algorithm from Ferrando et al. (2020) to perform full inference with other mechanisms. Ferrando et al. (2020)’s approach employs a parametric bootstrap to approximate the distribution of the coefficients’ estimator while accounting for the underlying data-generating distribution and the DP mechanism (see Algorithm 3 in Ferrando et al. (2020) for the method’s implementation). Although the original method uses the Laplace mechanism to achieve DP, we can employ this same technique with other mechanisms after making some simple adaptations. For that reason, we adapted Algorithm 3 to compare the method’s performance using the Analytic Gaussian mechanism from Balle and Wang (2018), the mechanism in Algorithm 2 from Sheffet (2019) (hereafter, the Wishart mechanism), and the mechanisms in Algorithm 2 from Wang et al. (2019) (hereafter, the Regularized Normal and Regularized Spherical Laplace mechanisms).

Ferrando et al. (2020)’s Algorithm 3 inputs are all functions of the noisy version of the summary statistics $S_H = S + H$, where $S = [X, Y]^t[X, Y]$ is the sufficient statistic for the linear model $Y = X\beta + \mathbf{e}$, Y is the vector representing the observations for the response, X is the design matrix, \mathbf{e} is the vector of independent and identically distributed normal errors, and the matrix H denotes the noise added to achieve differential privacy. Algorithm 3 also uses the parameter defining the corresponding DP mechanism, i.e., ϵ , δ , and the GS. Additionally, this algorithm assumes that the sample size n is known, which might not be true in many applied scenarios. Thus, we replace the unknown sample size with a noisy version of it. Note that when the model includes the intercept, a privatized version of the sample size is in the entry (1,1) of S_H . Otherwise, users will need to spend part of their privacy budget querying this quantity.

The implementation of Ferrando et al. (2020)’s algorithm also requires that S_H be positive

definite, which is not always guaranteed in practice. For this reason, we use a regularization whenever needed, i.e., $S_H^r = S_H - rI_{(p+1) \times (p+1)}$, such that r is equal to zero if S_H is definite positive or less than the minimum eigenvalue of S_H otherwise. Thus, S_H^r is guaranteed to be positive-definite. To find r for the Wishart mechanism, we use Remark (2) of [Sheffet \(2019\)](#). The remark contains an analytic expression to ensure that S_H^r is positive definite with probability greater than $1 - \delta$. For the Laplace and Analytic Gaussian mechanisms implementations, we follow a similar idea and define r , such that $P[H - rI_{p \times p} \text{ definite positive}] \approx p_0$ with a p_0 close to one. For the Regularized Normal and Regularized Spherical Laplace mechanisms, we define S_H^r as the already-regularized resulting matrix obtained from [Wang et al. \(2019\)](#)'s Algorithm 2.

Unfortunately, we only have an analytical expression for r for the Wishart mechanism. Hence, as proposed in [Peña and Barrientos \(2021\)](#), we employ simulations to approximate the distribution of the smallest eigenvalue of H and define r to be the p_0 -percentile of this distribution. Note that $P[S_H^r \text{ definite positive}] > P[H - rI_{p \times p} \text{ definite positive}]$. When defining r as before, r sometimes fails to make S_H^r positive definite, with small probability. Thus, we define r as equal to three times the minimum eigenvalue of S_H , which was chosen based on empirical testing. Our adaptation of [Ferrando et al. \(2020\)](#)'s algorithm is provided in the Supplemental Materials.

The effect of H on the inferences should diminish for large sample sizes. In fact, Theorem 1 in [Ferrando et al. \(2020\)](#) shows that $\hat{\beta}_H$ and $\hat{\beta}$ share the same asymptotic normal distribution. Their results also extend to mechanisms different to the Laplace mechanism, such as the Analytic Gaussian and Wishart. For this reason and to provide another method, we consider ignoring the noise added to S and plugging-in S_H in the formulas derived for linear regression models to compute point and interval regression coefficient estimates in the absence of noise.

Since the formulas for CIs also require that S be positive-definite, we use the regularized version of S_H , i.e., S_H^r . Therefore, we consider six methods, the Laplace mechanism, the BHM mechanism, and adaptations to the Analytic Gaussian, Wishart, Regularized Normal, and Regularized Spherical Laplace mechanisms, each using both of [Ferrando et al. \(2020\)](#)'s approaches except the BHM mechanism. We summarize the methods selected for evaluation in Table 1.

In addition to the methods based on [Ferrando et al. \(2020\)](#)'s approach, we consider the boot-

Table 1: Summary of the differentially private regression methods we tested for our tax use case studies. All methods return two types of confidence intervals (bootstrap-based and plug-in-based asymptotic) estimated using an adaptation from [Ferrando et al. \(2020\)](#), with the exception of BHM.

Method	Source
Analytic Gaussian mechanism	Balle and Wang (2018)
Laplace mechanism	Ferrando et al. (2020)
Regularized Normal mechanism	Algorithm 2 (Wang et al., 2019)
Regularized Spherical Laplace mechanism	Algorithm 2 (Wang et al., 2019)
Wishart mechanism	Algorithm 2 (Sheffet, 2019)
Browner-Honaker Method (BHM)	Browner and Honaker (2018)

strap approach described in Section 6.3 of [Browner and Honaker \(2018\)](#). We can directly apply this method when the summary statistic of interest is an average or a sum, such as S . Unfortunately, the strategy proposed by [Browner and Honaker \(2018\)](#) to compute CIs cannot be used in a straightforward manner for the CIs of regression coefficients.

However, since this method allows drawing multiple realizations of the noisy S (after splitting the privacy parameters (δ, ϵ)), we can approximate the sampling distribution of the regression coefficients relying on an asymptotic assumption. Specifically, let $\hat{\beta}_{BH,1}, \dots, \hat{\beta}_{BH,K}$ represent K sampled regression coefficients. Each of these coefficients are computed based on a realization of the noisy S using the equivalent of $(\delta/K, \epsilon/K)$ -differential privacy under [Browner and Honaker \(2018\)](#)'s approach. Thus, we approximate the sampling distribution of the regression coefficients by using a multivariate normal distribution, where the mean and covariance matrix correspond to the sample mean and covariance matrix based on $\hat{\beta}_{BH,1}, \dots, \hat{\beta}_{BH,K}$. We detail how we compute the sensitivity of S for the Laplace mechanism in the Supplemental Materials.

4 Use Case Studies and Results

We use the SOI 2012 PUF to evaluate the DP methods. SOI created this file based on a subsample of the confidential taxpayer data. The file contains 200 variables with 172,415 records and represents United States tax filers. The PUF has a few notable features, such as a stratified probability sample, observational weights, and many variables with skewed distributions or have values predominantly zero. In this study, we do not consider the weights, since no current DP regression methods exist to handle those. A description of the additional case study using the CPS ASEC data can be found

in the Supplemental Materials, and the code can be found online.²

4.1 Tax Policy Case Studies

We based our tests on the types of analyses that tax economists would normally query for the PUF data. For example, a tax expert would likely query one or more of the following:

- **Counting query:** How many tax returns have salary and wage income in excess of \$100,000?
- **Mean statistic query:** What are the means for the total and subsets of the total population?
- **Quantile statistic query:** What is the income threshold for the top 10 percent of earners?
- **Regression query:** What effects do marginal tax rates have on capital investment?

For regression analysis, we replicate results from [Feldstein et al. \(1980\)](#) on the 2012 PUF. The model estimates the relationship between “first dollar” marginal tax rates on capital gains and the realization of long-term capital gains divided by dividends holding constant if the filer is over age 65, the natural log of dividends, and the natural log of net Adjusted Gross Income (AGI). We limit the model to observations with at least \$12,513 in dividends, which is \$3,000 in 1980 dollars.

We calculate the first dollar marginal tax rates on long-term capital gains with a tax calculator, an algorithm that applies tax laws to microdata. The first dollar rate is calculated by running a tax calculator with zero long-term capital gains for every record and running a tax calculator with one penny of long-term capital gains for every record and then comparing the amount of taxes paid. We specifically used the Policy Simulation Library’s tax calculator ([Tax-Calculator, 2021](#)). The 2012 PUF is missing for about half of observations for age. This creates issues for replicating [Feldstein et al. \(1980\)](#), so we impute age by assigning observations with missing age to age 65+ if the record has non-zero total pensions and annuities received, non-zero pensions and annuities in AGI, or Social Security benefits. All other records with missing age are assigned to ages 0-64. This is crude but should be somewhat precise for records with more than \$12,513 in dividends. Furthermore, the estimates with and without noise will be flawed in the same way.

We use [Feldstein et al. \(1980\)](#), which is older, because most newer analyses use panel data or more complex methods, such as regression discontinuity design. For the former, the formal privacy

²GitHub repo website, <https://github.com/UrbanInstitute/formal-privacy-comp-appendix>

research is limited on how to handle this data type. For the latter, there are little to no formal privacy methods that apply to those regression models. We discuss these limitations further as part of our future work. More details on the histogram, means, and quantile analyses used in our evaluation are provided in the Supplementary Materials.

4.2 Assumptions

As with any practical application, we encountered additional challenges and had to make certain assumptions. We assumed there were no empty subsets and no survey weights. In the naive case, if someone using the validation server applied their analyses to an empty subset, then they would receive an error message. An error message would inform the user that there are few or no observations in the interested subset, violating the guarantees of DP. We acknowledge the reality of this issue for creating a validation server, but addressing this problem is beyond our scope.

To the best of our knowledge, there are little to no DP methods for handling survey weights for regression. Some methods exist for tabular statistics, but early testing showed that the GS became too large to be useful when applied to our data. Although we do not explore weights, we will consider them for later stages of the validation server and discuss this further in Section 6.

4.3 Utility Metrics

For evaluating the methods, we utilize various statistical measures. For all of the methods, we report the bias and root mean squared error (RMSE). For means, we also compute the CI overlap (CIO) value. This metric is commonly seen in synthetic data literature, which compares the regression CIs from the original and synthetic data to see how much the privacy algorithm affects regression inference. [Karr et al. \(2006\)](#) proposed the measure defined as:

$$CIO = 0.5 \left(\frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_o - l_o} + \frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_s - l_s} \right) \quad (8)$$

where u_o , l_o and u_s , l_s are the upper and lower bounds for the original and synthetic CIs respectively. The metric measures how much the CIs estimated the original and synthetic data overlap for a single estimate on average, where the maximum value is 1. The value is negative if the intervals do not overlap and grows more negative as they move further away. Note that a drawback to the IO

measure is the inability to distinguish whether the original or the synthetic data have a wider CI that encompasses the other interval. If one interval is wider but completely encompasses the other interval, the minimum value is 0.5 regardless of the width.

Given this issue, we calculate the ratio of the widths of the intervals (CIR) and also conducted two additional utility measures for our regression analyses. We calculate the number of times across the 100 repetitions that the coefficient signs of the original data match the DP coefficients and if the DP CIs included zero. These two measures allow us to more accurately assess the inferential differences (for hypothesis testing) between the original and the private outputs. In a practical setting, this evaluation tells us how the DP methods will impact the public policy decisions for the particular tax case study. Specifically, we evaluate whether the tax public policy outcome will change for each analysis using the noisy results.

5 Tabular and Summary Statistics Results

In this section, we present the results of the DP tabular, mean, and quantile methods. When testing our methods, we replicated all DP methods 100 times to assess variability and set $\epsilon = \{0.1, 0.5, 1, 5, 10, 15, 20\}$ and $\delta = \{10^{-3}, 10^{-7}, 10^{-10}\}$. We direct interested readers to our GitHub repo,³ which contains the code, data, and results from our study.

5.1 Histograms

We see a similar pattern in all three utility metrics between the two methods we tested. As expected, the Laplace mechanism outperforms the Gaussian mechanism, even for $\delta = 10^{-3}$. The maximum relative error is less than 1% for the Laplace mechanism at $\epsilon = 1$, and less than 1% for the Gaussian mechanism at $\epsilon = 5$. The mean cumulative sum error is less than 1% for the Laplace mechanisms when $\epsilon = 5$. We note that both methods are unbiased, and for $\epsilon \geq 5$ they both perform well enough that there is little difference between them. Overall, these results suggest the Laplace mechanism would perform well enough to enable researchers to query private histograms. More details on these results and figures can be found in the Supplementary Materials.

³GitHub repo website, <https://github.com/UrbanInstitute/formal-privacy-comp-appendix>

5.2 Quantiles

We found overall that both *AppIndExp* and *JointExp* offer high-quality quantiles, and they were generally preferable to *Smooth* for all but the highest values of ϵ . For $\epsilon < 5$, *Smooth* performs much worse than the other two methods, which is likely due to more extensive splitting of ϵ and δ required by the algorithm. We choose not to show the results of *Smooth* to improve the clarity of the plots. We find that *JointExp* performs better at estimating the zero-valued quantiles, though *AppIndExp* offers high-quality performance at most levels of ϵ and δ . Similarly, we see that *JointExp* has lower relative bias on average for the nonzero quantiles, but the difference between the two algorithms is very small. We also note that both algorithms show small but persistent bias, even for $\epsilon = 20$ (not shown), which suggests that the methods are not empirically unbiased. This may be an artifact of the method for sampling from the Exponential mechanism, which was implemented in the source code for these methods. For large values of ϵ , *Smooth* does improve and is unbiased unlike the other two methods. More details for these results can be found in the Supplementary Materials.

For a practical validation server implementation, *AppIndExp* and *JointExp* both appear sufficient to return accurate quantile estimates. The choice would likely depend on whether the system deploys pure DP or approximate DP. We do not recommend that a system use *Smooth* unless queries are made with very high levels of ϵ . Additionally, we find that *AppIndExp* returns more equally biased results for each quantile because they are drawn independently. On the other hand, *JointExp* can return biased results, which are more biased for some quantiles than others. In our application, the quantiles follow an exponential trend. *JointExp* returns more accurate results for the lowest and highest quantiles but returns less accurate results for those in-between. It may be preferable to choose one or the other algorithm, depending on the application.

5.3 Estimates of the Mean with Confidence Intervals

We found that all three methods for means are approximately unbiased. *NOISY-MAD* and *NOISY-VAR* perform similarly and both provide highly accurate statistics. BHM performs comparably to the other two methods only when $\delta = 10^{-3}$ and when $\epsilon < 1$. Otherwise, the other two clearly outperform BHM. We note that for higher ϵ the scale of the relative error is quite small for all methods (less than 1%) and would likely not effect the practical interpretation.

We found more variation for the CI measures than point estimate bias, which indicates the differing approaches to estimating uncertainty. Overall, we find *NOISYVAR* offers the best performance at all but the lowest ϵ . *NOISYMAD* performs well for $\epsilon = 0.1$, but as ϵ becomes larger, the width of the CIs produced by *NOISYMAD* shrinks and are consistently narrower than the confidential CI. These results would produce overly confident inference for researchers performing hypothesis testing. On the other hand, the average CIR and CIO for *NOISYVAR* moves closer to 1 as ϵ grows (as it does BHM but more slowly). More details on these results can be found in the Supplementary Materials. These results suggest that *NOISYVAR* is preferable overall, and it is capable of providing sufficiently accurate confidence intervals for higher levels of ϵ .

5.4 Regression Results

When testing our methods, we explored various values of ϵ and δ . We replicated all DP methods 100 times to assess variability. We set $\epsilon = \{0.5, 1, 5, 10, 15, 20\}$, $\delta = \{10^{-3}, 10^{-7}, 10^{-10}\}$, and a bootstrap size of 10 and 25 for the Brawner-Honaker method. We based these values from what is seen in other practical applications. Additionally, throughout this section, we limit figures to certain values of ϵ and δ depending on the range of the results in order to improve readability.

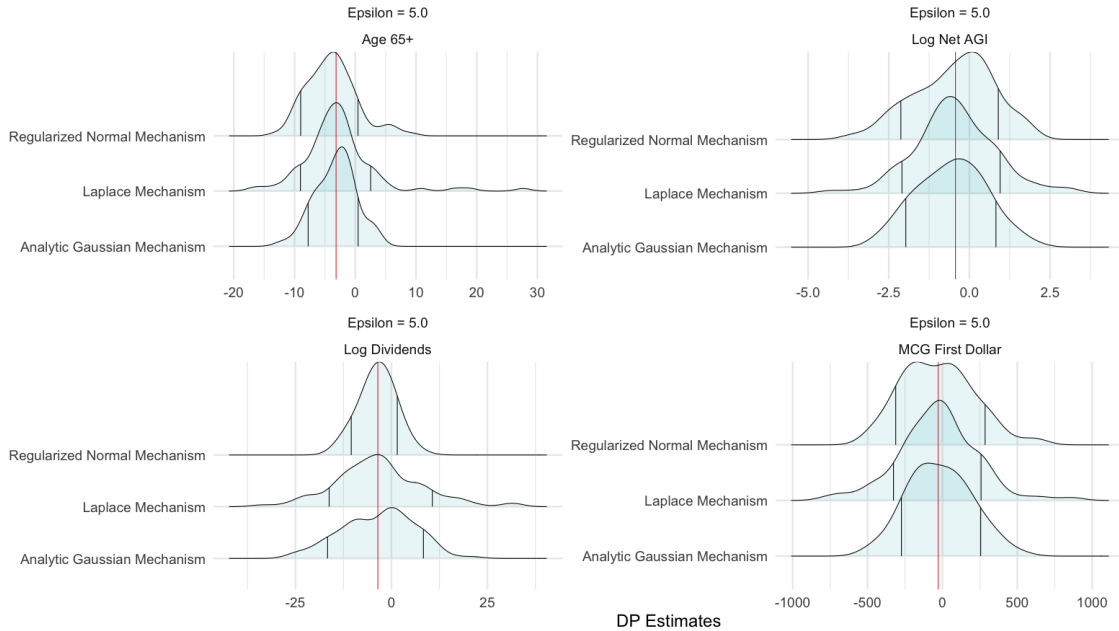


Figure 1: Distribution of Simulated DP Estimates for Regression Coefficients

We summarize our findings from the regression experiments, focusing on the methods that may

offer feasible results for a validation server. Unlike the tabular and summary statistics methods, the complexity of regression results makes it more difficult to assign the best performance to one method. For simplicity, we show results for the three best methods that have similar overall performance at $\epsilon = 5$ on the [Feldstein et al. \(1980\)](#) model described in Section 4.1. We provide the results for other ϵ and results on the CPS ASEC data in the Supplemental Materials.

We show findings for the four coefficients other than the intercept. We first consider the accuracy of the point estimates in absolute and relative terms. Figure 1 displays the distributions of simulated DP estimates for the coefficients, for three different methods at $\epsilon = 5$. The Regularized Spherical Laplace and Analytic Gaussian results shown use $\delta = 1e - 06$. The red lines indicate the confidential estimate, and the black lines indicate the 80% quantiles for the simulated DP estimates.

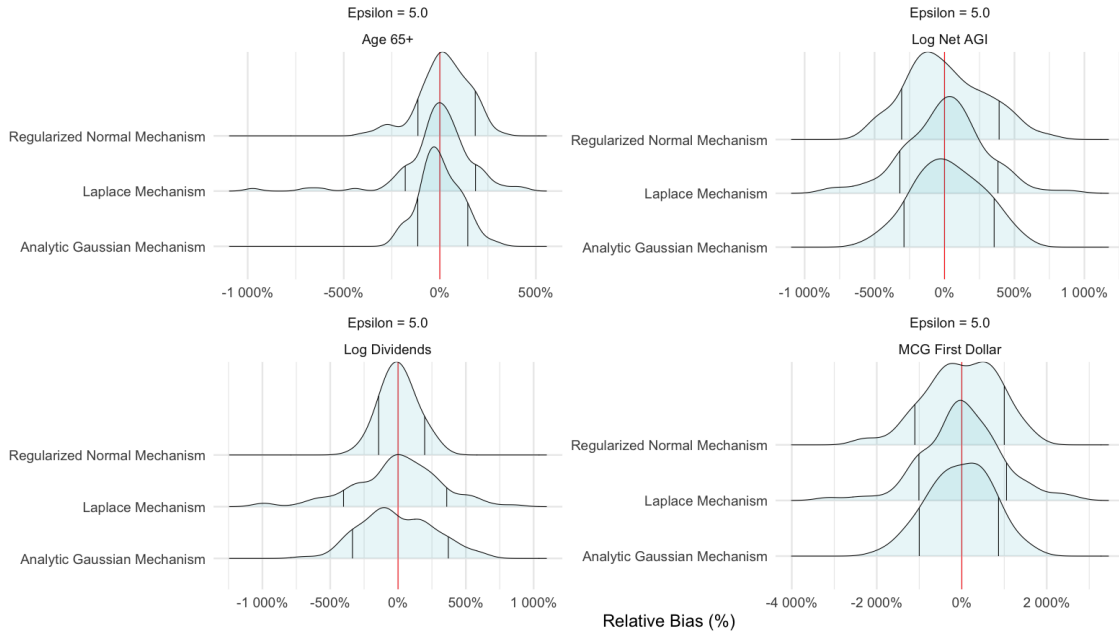


Figure 2: Distribution of Relative Bias for Regression Coefficients

Similarly, figure 2 shows the distributions of the bias as a percentage relative to the true estimate. This gives us an idea of the possible magnitude of bias. In this case, the red lines are at 0 for no bias. We see from these two figures that, while the distributions are generally centered at the confidential estimate, they are quite wide and include very large errors even at $\epsilon = 5$. We remind readers that these are the three methods which we found to have the best performance. While many queries returned for this model may contain satisfactory results, there is a reasonably high

probability that the query could return estimates that are very far from the confidential estimates.

We also evaluate the noisy CIs compared to the original CIs, because we are interested in evaluating uncertainty of the regression estimates. As described in Section 3.4, we utilize two different approaches for estimating CIs; bootstrap and plug-in asymptotic approaches. The former we expect to have wider CIs that account for the variability introduced by the noise mechanism. The asymptotic approaches, on the other hand, produce intervals with lengths more similar to the length of the confidential estimates because they assume large sample sizes that minimize the noise from the privacy mechanisms. Because both of these interval estimates can be output from the algorithms without spending additional budget, a validation server could return both “for free.” We thus evaluate the different estimates of uncertainty that these intervals provide.

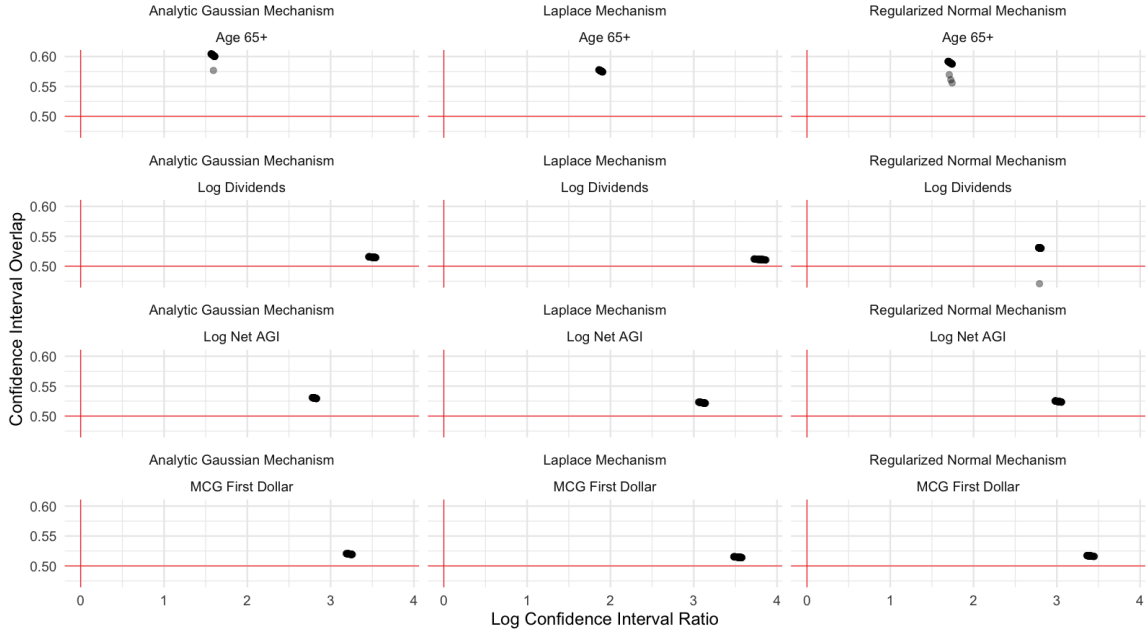


Figure 3: *Confidence Interval Overlap vs. Confidence Interval Ratios for each regression coefficient using the bootstrap variability estimates. The vertical red lines indicate a CIR of 1 and the horizontal red lines indicate a CIO of 0.5. $\epsilon = 5$.*

For figures 3 and 4, we show the distributions of CIO and CIR values. We plot these metrics against each other because taken together they greatly aid the interpretation of the results. The red lines plotted at $CIO = 0.5$ and $CIR = 1$ form four quadrants on the charts. We summarize the interpretation of each quadrant as: top left quadrant indicates noisy CIs that are more narrow than the confidential CI but mostly contained within the confidential CI; top right quadrant indicates

noisy CIs that are wider than the confidential CI but mostly encompass the confidential CI; bottom left quadrant indicates noisy CIs that are both more narrow than the confidential CI and biased away from the confidential point estimate; and bottom right quadrant indicates noisy CIs that are both wider than the confidential CIs and biased away from the confidential point estimate.

Figure 3 shows the results for the bootstrap CIs. As a reminder, the results shown used $\epsilon = 5$. The plots indicate that in general the noisy intervals completely cover the confidential intervals but are much wider, as much as 50 times wider for some coefficients. We see that the categorical predictor has the narrowest intervals relative to the confidential, followed by the logged predictors, and finally the marginal tax rates predictor. This suggests that in order to account for the noise from the mechanisms, a high level of uncertainty must be added to the estimates. It also suggests that the noise varies substantially based on the predictor type and distribution. The marginal tax rates predictor is both bi-modal and heavily skewed, which likely means that the noise added is proportional to values greater than the majority of the observed values.

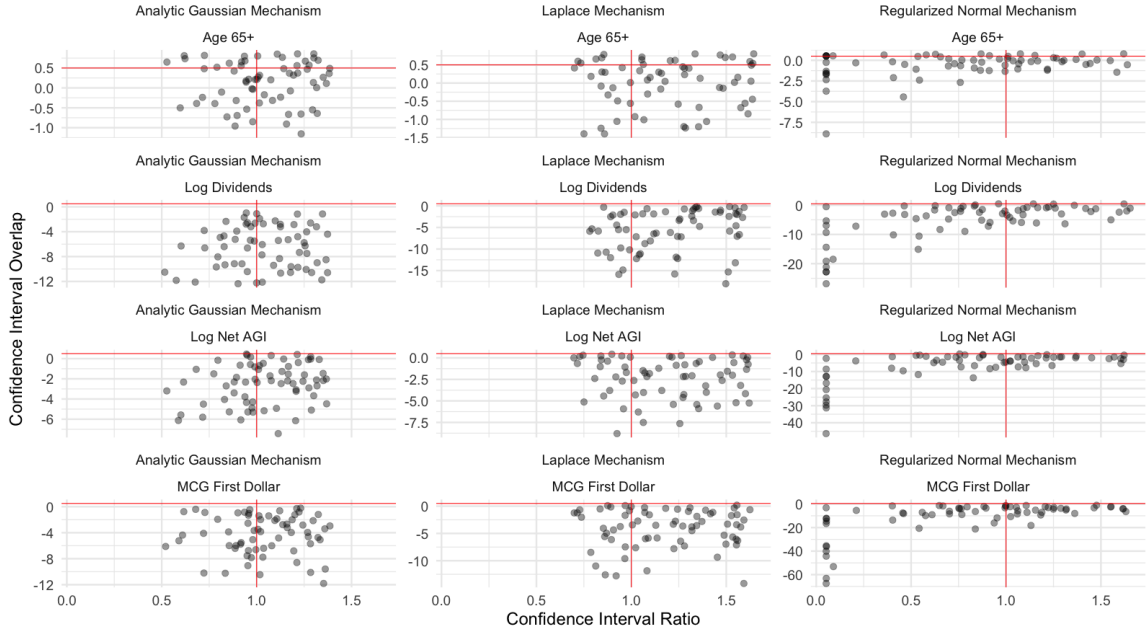


Figure 4: Confidence Interval Overlap vs. Confidence Interval Ratios for each regression coefficient using the asymptotic variability estimates. The vertical red lines indicate a CIR of 1 and the horizontal red lines indicate a CIO of 0.5. $\epsilon = 5$.

Alternatively, figure 4 shows very different results for the asymptotic CI estimates. As expected, the CIR are centered around 1, but the majority of the results have negative CIO values. This

indicates that most noisy CIs here do not overlap at all with the confidential intervals. Given the wide distribution of estimates seen in figure 1, this is understandable. CIs which do not take into account the noise added will frequently miss the confidential estimates completely.

Lastly, we look at regression inferences from a purely hypothesis testing perspective by measuring how frequently the noisy results match the sign and significance of the confidential estimates. In this case, we assume researchers are primarily interested in understanding whether a predictor has a positive or negative relationship and whether that relationship is statistically significant. We also assume they are less concerned about the exact magnitude of the estimate.

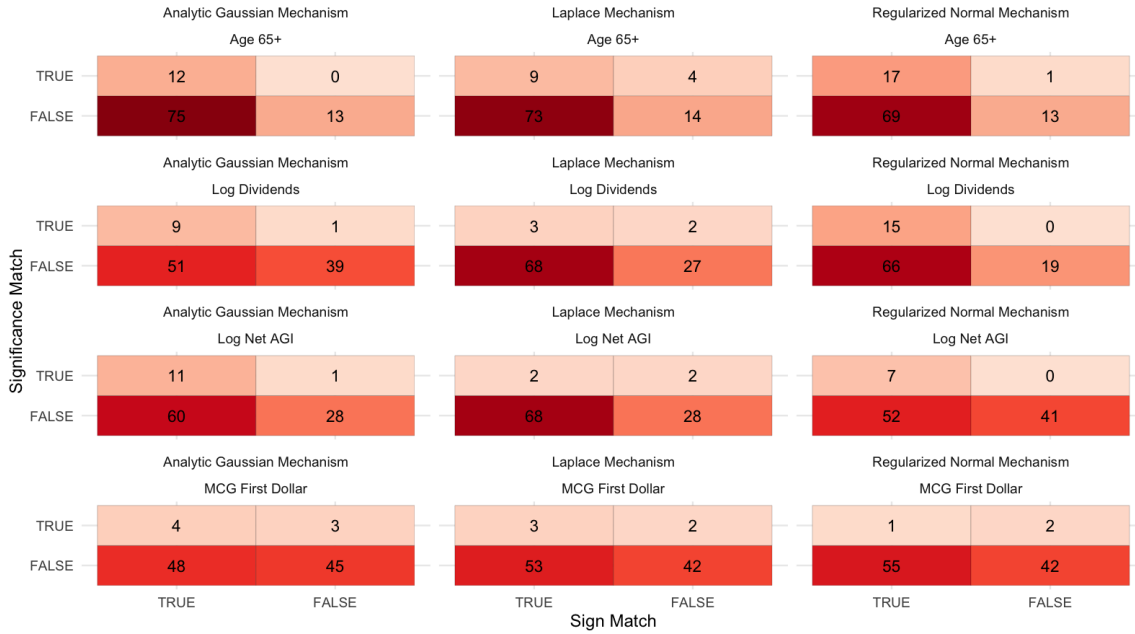


Figure 5: Confusion matrix showing percentages of sign and significance matches for each regression coefficient using the bootstrap confidence intervals. The upper left hand box indicates results that matched both. $\epsilon = 5$.

We look at these results for both the bootstrap and asymptotic interval estimates again, and the results are shown in figures 5 and 6 respectively. These results perhaps paint the most hopeful picture of any of the results for regression models, but they still have significant limitations. Both approaches achieve relatively high sign match, which suggests that, though biased, the results will more often point in the right direction. The significance match varies quite a lot by the uncertainty method used. All regression predictors were significant in the confidential models. Most of the time, we see that the bootstrap estimates are not significant because of the very wide intervals. On the

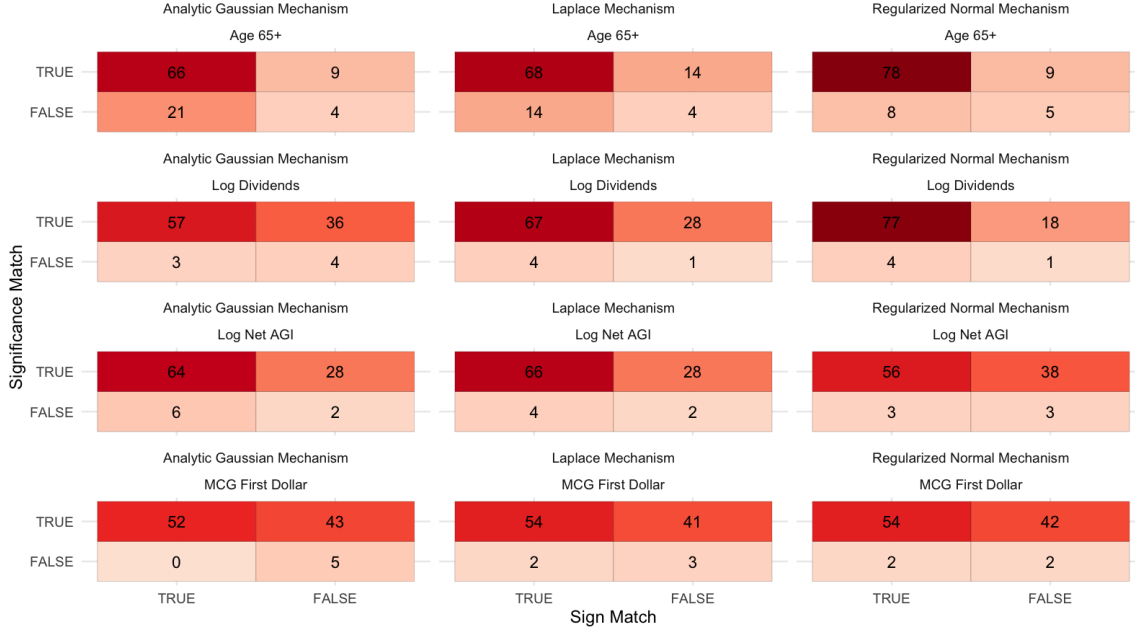


Figure 6: Confusion matrix showing percentages of sign and significance matches for each regression coefficient using the asymptotic confidence intervals. The upper left hand box indicates results that matched both. $\epsilon = 5$.

other hand, the asymptotic approaches all match significance at high levels, and they match both sign and significance for the majority of the results. The drawback for the asymptotic approach is that they have a high percentage of results that match the significance but not the sign, which could be seen as the most undesirable result. The asymptotic is more accurate for the significance because it will tend to produce narrower CIs that do not include zero. Since all the coefficients in the case study are significant, the asymptotic plug-in strategy matches significance frequently. If we had non-significant coefficients, the asymptotic would have likely performed worse. In summary, the bootstrap approach is neither likely to result in correct or incorrect inference (of a significant result), and the asymptotic can be seen as a high-risk, high-reward approach.

6 Conclusions from the Case Study

In this paper, we surveyed and tested the feasibility of the latest DP methods for summary statistics and regression analyses on real tax data. To the best of our knowledge, this is the first comprehensive evaluation of these DP methods for practical applications within a validation server framework for real-world data. We found that DP algorithms for summary statistics performed well if the pri-

vacy loss budget is larger than 1, whereas methods for regression analyses still need improvement when it comes to performing full inference. Practical applications would likely require either larger sample sizes or allocating more ϵ on every query in order to return estimates with satisfactory levels of uncertainty. Based on our study results, we identified a few challenges and avenues for future work.

6.1 Challenges

Through our study, we found that existing DP regression analyses were limited in applicability and often added more noise than needed to protect privacy. For example, suppose there are four coefficients in the regression model. The tested DP methods add noise to the design matrix, which has 10 sufficient statistics. The added noise is multiplicative rather than additive, resulting in highly noisy estimates that were of little practical use. Other DP methods we reviewed in Section 3.4 could possibly perform better in other evaluations, but they did not meet our inclusion criteria, such as not reporting standard errors. Similar to some of the mean and quantile methods, more research is needed to develop methods that are robust to other data types. Many of the methods we reviewed were limited to particular applications or had unrealistic assumptions about the data.

Besides the methodological issues, we encountered challenges with coding the various algorithms. We sometimes discovered errors in pseudocode from a manuscript and from code we collected from GitHub or from the author(s). Overall, these were minor issues, and we informed the author(s) of bugs. However, obtaining code for applications was more problematic. Though we do not expect privacy researchers to provide production-ready code, we often discovered the research code to be messy, hard to read, and difficult to alter for our use cases. This situation is still preferable to not having any code or hand-coding the method based on the paper. We encountered this problem a few times, which prevented us from implementing some approaches. We reached out to author(s) when no code was available. If the author(s) did not respond, we attempted to code the methods ourselves. But, in some cases, the manuscript did not provide enough information for us to implement the method and was thus excluded from the feasibility study.

These issues emphasize the importance of open-source code to facilitate wider use and acceptance of DP algorithms in practice. For example, OpenDP from Harvard University and SmartNoise from Microsoft are developing a suite of open-source software tools to implement DP methods.

These platforms are still under development, which is why we did not use their code for the feasibility study, but they may offer improved solutions in the future.

6.2 Future Work

One vital area for improvement is developing DP algorithms for data that are not Gaussian. Many of the methods we tested performed well when originally proposed, because the authors tested on well-behaved or normally distributed data. Real data are often skewed, such as the 2012 PUF and CPS ASEC data, which resulted in these same methods performing poorly.

Another area for improvement is developing DP algorithms with a focus on estimating the uncertainty of the estimates. Many data privacy experts create DP regression analyses to output accurate predictions. But, as seen in Section 4, these methods performed poorly either by returning very large confidence intervals or by not reporting the standard error at all. This appears to be a significant gap in the DP literature that must be addressed.

With these potential research areas in mind, our future work developing a validation server will focus on improving formal privacy methods for OLS regression coefficient and interval estimates and then expand that research to address other important economic use cases. Future areas for work indicated by tax experts are the incorporation of survey weights or regression discontinuity and regression kink designs. We are not aware of any research on formal methods to protect privacy in regression discontinuity and regression kink designs. We hope this study provides the data privacy and confidentiality community with a better understanding of the capabilities, limitations, and challenges of current DP methods for summary statistics and regression analyses.

Acknowledgments

This research was funded by the Alfred P. Sloan Foundation [G-2020-14024] and National Science Foundation National Center for Science and Engineering Statistics [49100420C002].

We would like to thank our collaborators at the Statistics of Income Division, especially Barry Johnson and Victoria Bryant, for their amazing support for this project. We also thank our stellar validation server project team, consisting of Leonard Burman, John Czajka, Surachai Khitatrakun, Graham MacDonald, Rob McClelland, Silke Taylor, Kyle Ueyama, Doug Wissoker, and Noah

Zwiefel. Thank you to Gabriel Morrison for reviewing our code. Finally, we thank our advisory board for their invaluable advice and support. The members are John Abowd, Jim Cilke, Jason DeBacker, Nada Eissa, Rick Evans, Dan Feenberg, Max Ghenis, Nick Hart, Matt Jensen, Barry Johnson, Ithai Lurie, Ashwin Machanavajjhala, Shelly Martinez, Robert Moffitt, Amy O'Hara, Jerry Reiter, Emmanuel Saez, Wade Shen, Aleksandra Slavković, Salil Vadhan, and Lars Vilhuber.

Supplemental Materials for

“A Feasibility Study of Differentially Private Summary Statistics and Regression Analyses for Administrative Tax Data”

This file contains the Supplemental Materials to accompany the paper “A Feasibility Study of DP Summary Statistics and Regression Analyses for Administrative Tax Data.”

7 Tabular and Summary Statistics Differentially Private Algorithms

For this section, we review the DP tabular, mean, and quantile algorithms that we considered and which we selected for the feasibility study.

7.1 Candidate and Selected Methods

Table 2: Summary of the DP tabular methods we reviewed in Section 3.1.

<i>Tabular Statistics</i>			
Method	Privacy Definition	Off-the-Shelf vs. Hand-Coding	Selected for Case Study
Laplace mechanism	ϵ -DP	off-the-shelf via R and Python code on GitHub	Yes
Gaussian mechanism	(ϵ, δ) -DP	off-the-shelf via R and Python code on GitHub	Yes

Table 3: Summary of the DP quantile statistics reviewed in Section 3.2.

<i>Quantile Statistics</i>			
Method	Privacy Definition	Off-the-Shelf vs. Hand-Coding	Selected or Not Selected for Case Study
<i>AppIndExp</i> (Gillenwater et al., 2021; Smith, 2011)	(ϵ, δ) -DP	off-the-shelf via Python code on GitHub	Yes
<i>JointExp</i> (Gillenwater et al., 2021)	ϵ -DP	off-the-shelf via Python code on GitHub	Yes
<i>Smooth</i> (Nissim et al., 2007)	(ϵ, δ) -DP	off-the-shelf via Python code on GitHub	Yes
<i>Concentrated Smooth</i> (Gillenwater et al., 2021)	CDP	off-the-shelf via Python code on GitHub	No, requires fine tuning which is not realistic for our application
Propose-Test-Release (Dwork and Lei, 2009)	(ϵ, δ) -DP	No code publicly available	No, requires fine tuning which is not realistic for our application
Tzamos et al. (2020)	ϵ -DP	No code publicly available	No, requires fine tuning which is not realistic for our application

Table 4: Summary of the DP mean confidence interval methods we reviewed in Section 3.3.

<i>Confidence Intervals for the Mean</i>			
Method	Privacy Definition	Off-the-Shelf vs. Hand-Coding	Selected or Not Selected for Case Study
Brawner and Honaker (2018)	$\text{zCDP} \leftrightarrow (\epsilon, \delta)\text{-DP}$	off-the-shelf via R code on GitHub	Yes
D’Orazio et al. (2015)	$\epsilon\text{-DP}$	off-the-shelf via R code on GitHub	No, requires a priori bounds set on standard deviation
Karwa and Vadhan (2017)	$(\epsilon, \delta)\text{-DP}$	off-the-shelf via R code on GitHub	No, requires a priori bounds set on standard deviation
<i>COINPRESS</i> (Biswas et al., 2020)	zCDP	off-the-shelf via R code on GitHub	No, requires a priori bounds set on variance/covariance
<i>NOISYMAD</i> (Du et al., 2020)	$\epsilon\text{-DP}$	off-the-shelf via R code on GitHub	Yes
<i>NOISYVAR</i> (Du et al., 2020)	$\epsilon\text{-DP}$	off-the-shelf via R code on GitHub	Yes
<i>CENQ</i> (Du et al., 2020)	$\epsilon\text{-DP}$	off-the-shelf via R code on GitHub	Yes, for non-skewed data
<i>MOD</i> (Du et al., 2020)	$\epsilon\text{-DP}$	off-the-shelf via R code on GitHub	Yes, for non-skewed data
<i>SYMQ</i> (Du et al., 2020)	$\epsilon\text{-DP}$	off-the-shelf via R code on GitHub	Yes, for non-skewed data

8 Tax Policy Case Studies

We base our histogram analysis on a report by [Mortenson and Whitten \(2020\)](#), where the authors calculate bunching estimators on detailed histograms of univariate distributions. Their histogram presents earned income in \$200-wide bins for single taxpayers with two children in 2014 from \$0 to \$30,000. Their goal is to identify bunching around the first earned income tax credit kink point. We produce this histogram on the 2012 PUF data. As part of preprocessing step for the histogram, we calculated the earned income for filers with no dependent-status indicator who file as single or head of household. Dependents is the sum of exemptions for children living at home, exemptions for children living away from home, and exemptions for other dependents. Earned income is wage and salary income, positive net business income, and positive net farm income.

9 Tabular and Summary Statistics Results

In this section, we present figures for the results discussed in the main report.

9.1 Histograms

We tested the two fundamental mechanisms, Laplace and Gaussian, for providing tax data histograms. We also tested making separate queries for each bin of the histogram, but results were worse than the multivariate versions. We calculate the RMSE for the counts within each bin, and the results are shown in [Figure 7](#).

In addition to RMSE, we utilize a couple other utility metrics which allow us to interpret the error distributions in such a way that relates to how a tax researcher might use the queried histograms. In particular, discerning cut points or jumps in the distribution is important, which rely on understanding the cumulative distribution function. We measure utility on the CDF in two interpretable ways. First, we measure the maximum error of over all cumulative sums relative to the sum over the whole histogram. This tells the maximum percentage error for someone trying to determine the what percentage of individuals fall below or above a certain cutoff. Second, we measure the mean error over all cumulative sums for the histogram, which tells us the average error across all adjacent subsets of the histogram. Results are shown in [figures 8 and 9](#) respectively.

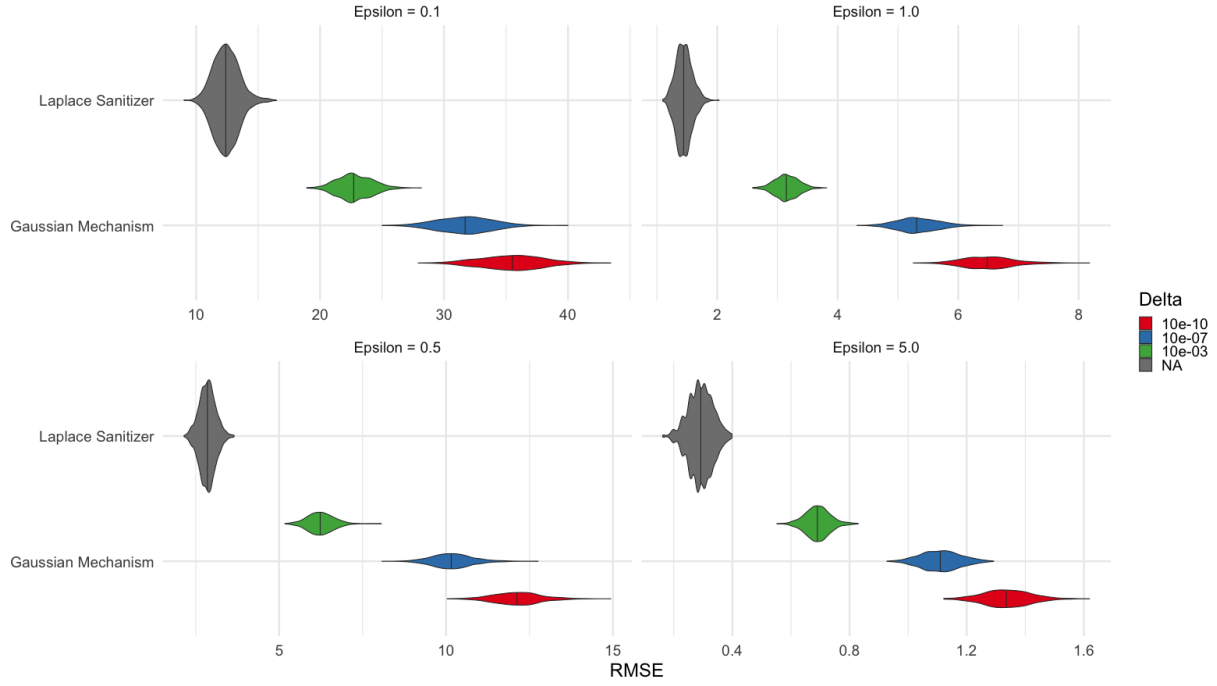


Figure 7: Income Histogram Results - RMSE

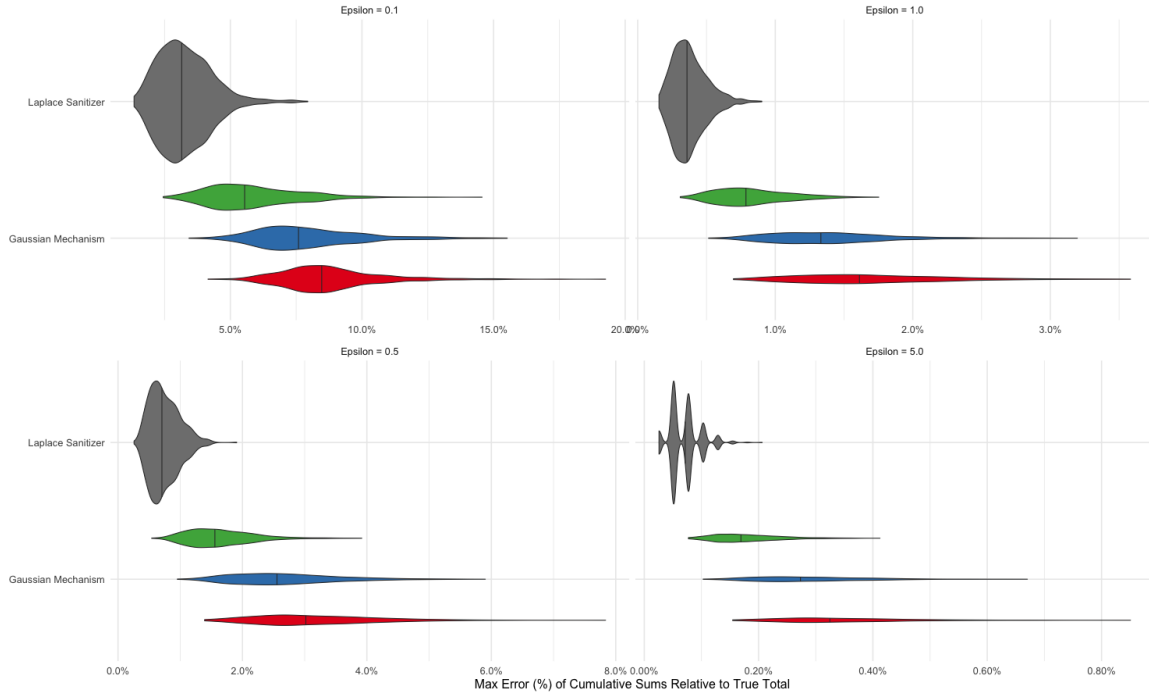


Figure 8: Income Histogram Results - Max Relative Error

9.2 Quantiles

Users may also query certain quantiles as a way to understand the distribution of data, such as income, rather than querying a histogram or mean value. We tested three different methods for

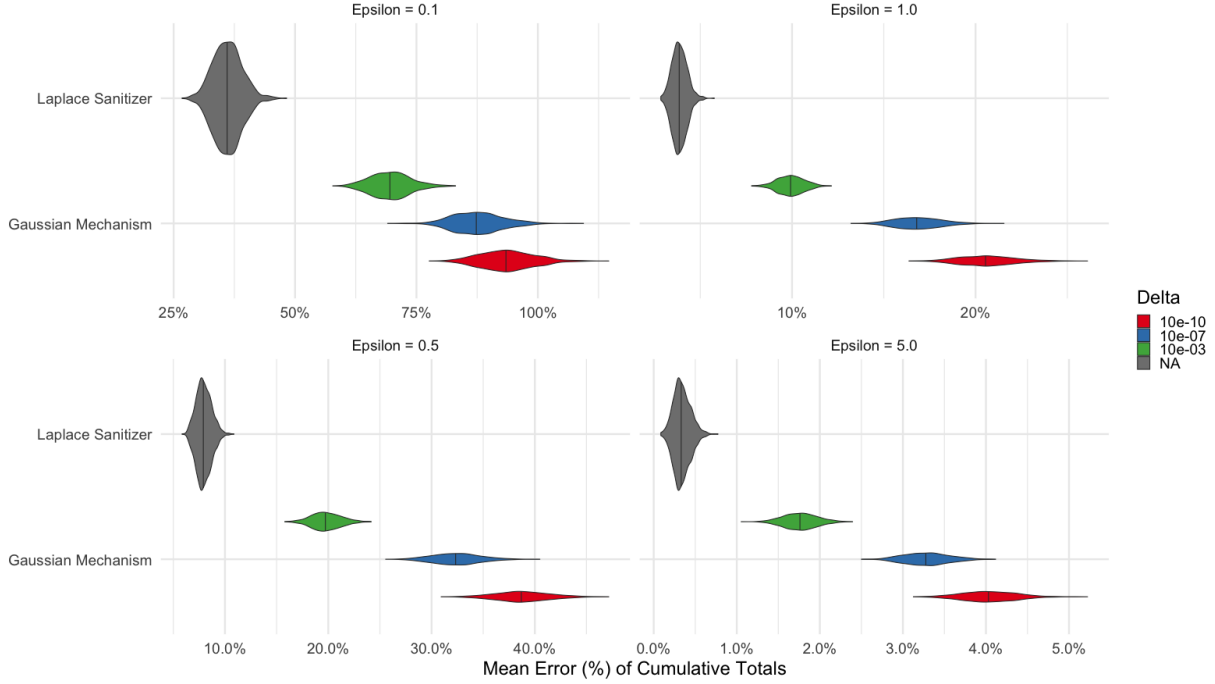


Figure 9: Income Histogram Results - Mean Cumulative Sums Error

producing multiple quantiles, which is likely to be of interest for those making queries. We show the results in figures 10 and 11. 2 of the 9 estimated quantiles have confidential values equal to 0. For readability, we first show the mean absolute bias for those two quantiles, and then show the mean relative bias across the other seven quantiles.

9.3 Estimates of the Mean and Confidence Intervals

We tested the accuracy of estimates for mean income and the accuracy of confidence intervals for the statistic. We compared three methods capable of computing a confidence interval for the mean with only a priori assumptions on the data bounds. Three additional methods listed in table 4 were not tested due to the heavy skew of the data. We tested them on a subset of the CPS data, which was not skewed, but still found poor results. This indicated that they would only provide unbiased results for nearly perfectly Gaussian data. See Appendix 12 for more details. Figure 12 shows the results for the mean bias of the point estimate.

Figure 13 shows the simulated distributions of confidence interval overlap (CIO) and confidence interval ratio (CIR) values. We plot these two metrics against each other because taken together they greatly aid the interpretation of the results. The red lines plotted at $CIO = 0.5$ and $CIR = 1$

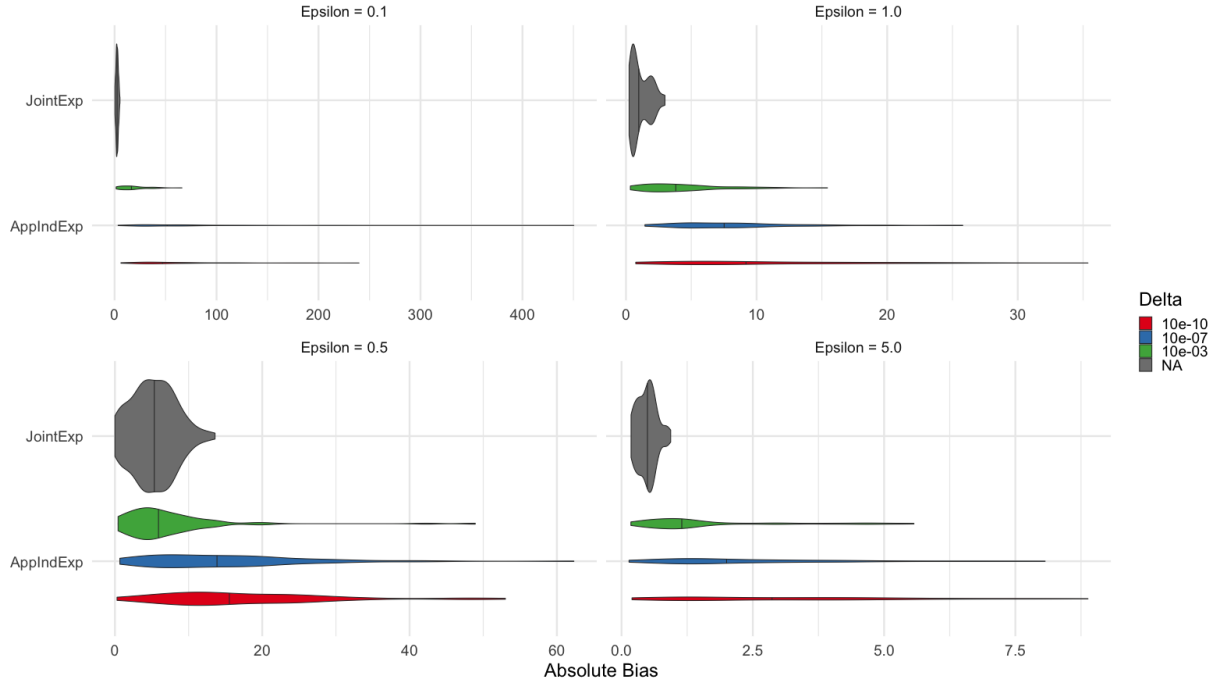


Figure 10: Quantile Absolute Bias Results, Only Quantiles with Confidential Values = 0

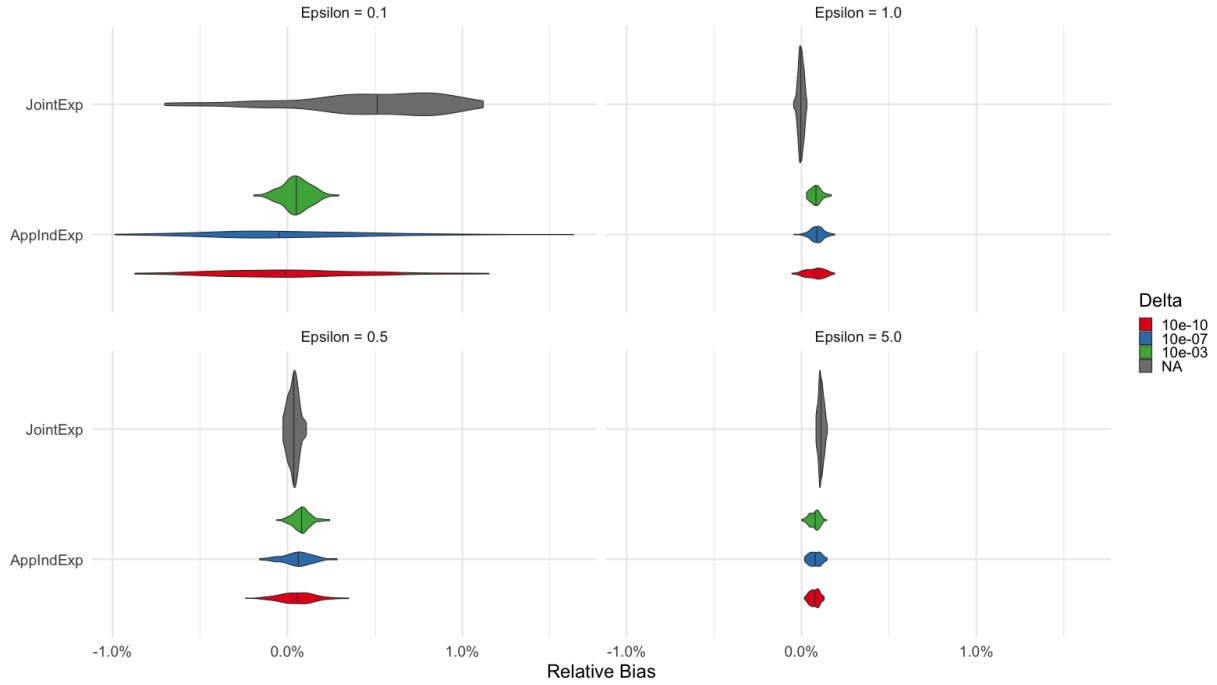


Figure 11: Quantile Relative Bias Results, Only Quantiles with Confidential Values > 0

form four quadrants on the charts. We summarize the interpretation of each quadrant as follows:

- **Top left quadrant:** indicates noisy CIs that are more narrow than the confidential CI but

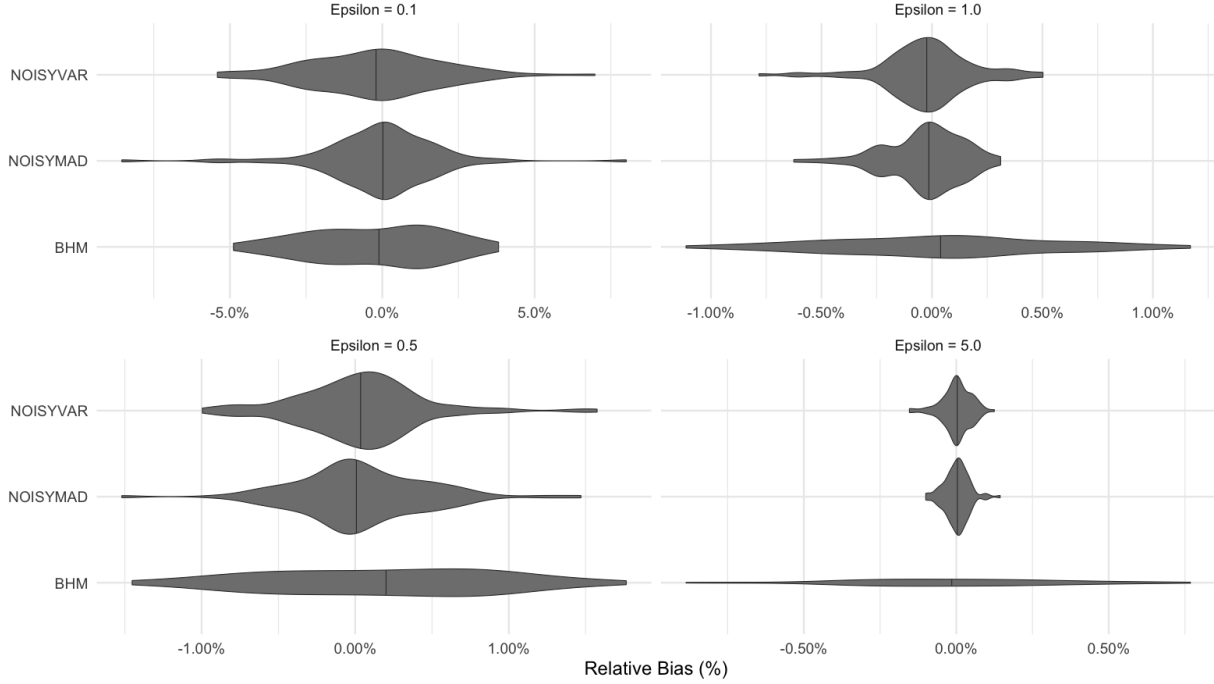


Figure 12: Mean Income Relative Bias. BHM Results Only Shown for $\delta = 0.01$.

mostly contained within the confidential CI.

- **Top right quadrant:** indicates noisy CIs that are wider than the confidential CI but mostly encompass the confidential CI.
- **Bottom left quadrant:** indicates noisy CIs that are both more narrow than the confidential CI and biased away from the confidential point estimate.
- **Bottom right quadrant:** indicates noisy CIs that are both wider than the confidential CIs and biased away from the confidential point estimate.

10 Extended Review of Differentially Private Regression Analyses

As stated in the main text, DP methods for regression can be classified according to the outputs they produce: (1) point estimates only, (2) point estimates and interval estimates, and (3) other outputs related to regression analysis, such as diagnostic plots. In this section, we cover categories (1) and (3).

Most DP algorithms fall into the first category, returning only noisy point estimates. Methods in this category frequently rely on objective-perturbation-based approaches, such as the functional

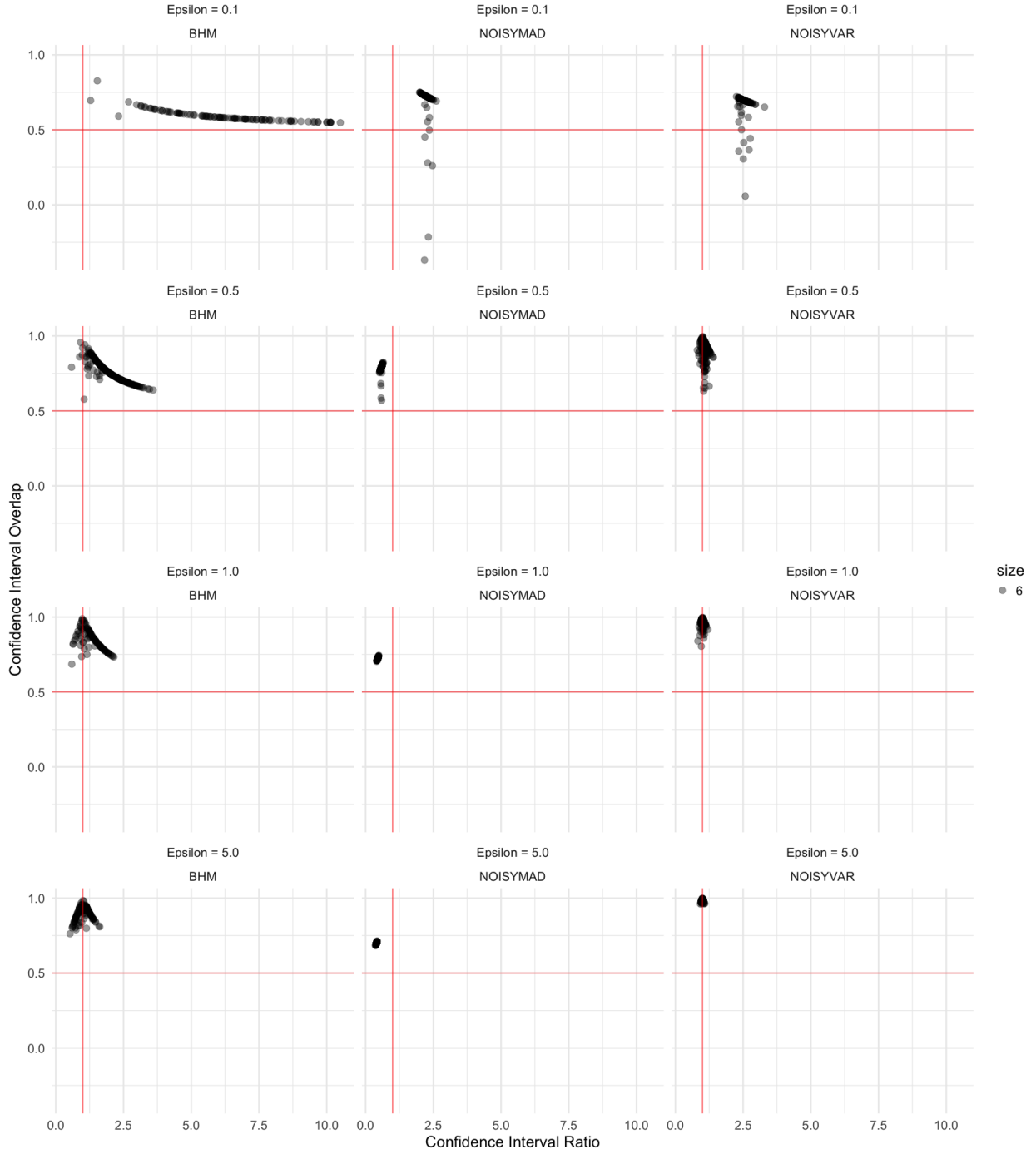


Figure 13: Mean Income Confidence Intervals Results. BHM Results Only Shown for $\delta = 0.01$.

mechanism (Chaudhuri et al., 2011; Fang et al., 2019; Gong et al., 2019; Zhang et al., 2012). These approaches provide DP coefficients estimates by maximizing a perturbed version of the objective function. We can obtain the perturbed versions of the objective function by either adding noise to

the function or using a polynomial representation of the function, where we add noise polynomial coefficients.

Other DP point estimate methods borrow ideas from robust statistics. For example, [Avella-Medina \(2020\)](#) proposes (ϵ, δ) -differential private methods to obtain robust estimators for various problems, including regression analysis. [Chen et al. \(2020\)](#) developed ϵ -DP methods for median regression, which can be seen as a robust version of ordinary least squares based regression. Finally, some DP point estimate approaches add noise to sufficient statistics ([Wang, 2018](#)). For a list of DP methods for simple linear regression, we refer the reader to [Alabi et al. \(2020\)](#).

Next, we review recent contributions that release outputs distinct from point and interval estimates. Even though these methods are beyond the current scope of the validation server, they provide key information for regression analysis and could be fruitful additions to future versions of the validation server.

[Barrientos et al. \(2019\)](#) proposed an ϵ -DP method to perform hypothesis testing for single coefficients. This approach has the advantage of not requiring the response and predictors to be bounded. However, the privacy loss budget may be costly when performing hypothesis testing for multiple coefficients using this method. The Bayesian procedure by [Amitai and Reiter \(2018\)](#) has similar capabilities and limitations, since it targets specific summaries of the posterior distributions of the regression coefficients, such as tail probabilities.

Residual analysis is another important task in regression and a crucial tool for model validation when finding solutions to problems, such as heterogeneity of errors and lack of linearity. [Chen et al. \(2016\)](#) describe ϵ -DP methods that release a private version of the residuals versus fitted values plot. Model selection is also key for regression analysis, and [Lei et al. \(2016\)](#) has developed an (ϵ, δ) -DP algorithm for this purpose. Other contributions related to regression analysis involve algorithms developed for regularized regression, such as Lasso ([Dandekar et al., 2018](#); [Talwar et al., 2015](#)).

10.1 Details on Differentially Private Bootstrap Regression

[Ferrando et al. \(2020\)](#) proposed a DP bootstrap-based algorithm that adds noise to the sufficient statistic for linear model $Y = X\beta + \mathbf{e}$, where Y is the vector representing the observations for the

response, X is the design matrix, and \mathbf{e} is the vector of independent and identically distributed normal errors. We assume that the reported noisy statistic is of the form

$$S_H^r = S + H - rI_{(p+1) \times (p+1)}$$

where $S = [X, Y]^t[X, Y]$ is the sufficient statistic for linear model $Y = X\beta + \mathbf{e}$, the matrix H denotes the noise added to achieve differential privacy using either the Laplace, Analytic Gaussian, or Wishart mechanism, and the r is defined as discussed in Section 3.4.2. For the Normal and Spherical Laplace mechanisms, we define S_H^r as the already-regularized resulting matrix obtained from Wang et al. (2019)’s Algorithm 2. We denote P_H as the corresponding probability distribution of H under a given mechanism. To define H for the Normal and Spherical Laplace mechanisms, we use a symmetric version of the added noise as in Wang et al. (2019)’s Algorithm 2, i.e., we define $H = (\tilde{H} + \tilde{H}^t)/2$ where the entries of \tilde{H} are drawn from the corresponding normal or spherical Laplace distribution. The up-to-date version of the algorithm proposed by Ferrando et al. (2020) assumes that r is equal to zero and only considers the Laplace mechanism. This algorithm also assumes that the sample size n is known. This is something we cannot assume and, for this reason, we replace the sample size by a DP version of it. Notice that when the intercept is included in the model and represented by the first column of X , a privatized version of the sample size is available at the entry (1,1) of S_H^r . If the intercept is not included, users will need to spend part of their privacy budget querying this quantity. The algorithm below summarizes the employed algorithm after modifications and adaptations.

10.2 Sensitivity Calculations

To compute the sensitivity of S for the Laplace mechanism described in Section 3.4.2, we assume that the response and predictors are bounded—a common assumption in differential privacy. Without loss of generality, we assume that the response and predictors take values $[0, 1]$. Because S is a symmetric matrix and under the previous assumption, its sensitivity is upper-bounded by the number of entries in and above the diagonal, i.e., $(p+1)(p+2)/2$, where p is the number of regression coefficients. While using the upper bound $(p+1)(p+2)/2$ as sensitivity is a valid strategy, it is particularly inefficient in the presence of categorical predictors. To improve this upper bound,

Algorithm 1 DP bootstrap-based regression

Input:

S_H^r : regularized noisy sufficient statistic P_H : probability distribution of H
 B : number of bootstrap samples p : number of regression coefficients

```
1:  $\hat{n} \leftarrow S_H^r[1, 1]$      $\triangleright$  assuming the intercept is part of the model
2:  $\widehat{X^T X} \leftarrow S_H^r[1:p, 1:p]$      $\triangleright$  submatrix of  $S_H^r$  with  $1:p = (1, \dots, p)$ .
3:  $\widehat{X^T Y} \leftarrow S_H^r[1:p, p+1]$ 
4:  $\widehat{Y^T Y} \leftarrow S_H^r[p+1, p+1]$ 
5:  $\hat{\beta} \leftarrow (\widehat{X^T X})^{-1} \widehat{X^T Y}$ 
6:  $\hat{\sigma}^2 \leftarrow (\widehat{Y^T Y} - (\widehat{X^T Y})^T (\widehat{X^T X})^{-1} \widehat{X^T Y}) / (\hat{n} - p - 1)$ 
7: for  $b$  in  $\{1, \dots, B\}$  do
8:   Sample  $h$  from  $P_H$ 
9:   Sample  $\widetilde{X^T \mathbf{u}}$  from  $N(0, \hat{n} \hat{\sigma}^2 \widehat{X^T X})$ 
10:   $\tilde{\beta}_b \leftarrow (\widehat{X^T X})^{-1} (\widehat{X^T X} - h[1:p, 1:p]) \hat{\beta} + (\widehat{X^T X})^{-1} (\widetilde{X^T \mathbf{u}} + h[1:p, p+1])$ 
return:  $\tilde{\beta}_1, \dots, \tilde{\beta}_B$ 
```

we implement some strategies when categorical predictors are part of the analysis. We assume that all categorical predictors are included in the model as dummy variables and fixing one level as the reference. The implemented strategies are listed below:

1. One or more entries in the diagonal of S are identical to entries in the off-diagonal. Hence, we only need to add noise to the unique entries.
2. If the categorical predictor has more than two levels, then some entries of S will be exactly zero, eliminating the need to add noise to those entries.
3. If the categorical predictor has more than two levels and the model has an intercept, then we multiply the set of the corresponding dummy variables by the column of ones in X (representing the intercept) results in a vector that counts the number of ones in the dummy variables. The vector is equal to a contingency table that counts the number of observations at each level, leaving out the reference one. Although this vector has a dimension greater than one, its sensitivity is equal to one. Hence, we take advantage of this fact to reduce the sensitivity of S .
4. If the categorical predictor has more than two levels and the model has a numeric predictor, then we multiply the set of corresponding dummy variables by the column in X representing

the numerical predictor results in a vector of partial sums. Each entry of this vector is equal to the summation of numerical predictor values across the observations that share the same level as the categorical predictor. Similar to the previous item, while this vector has a dimension greater than one, its sensitivity is equal to one. Hence, we take advantage of this fact to reduce the sensitivity of S .

We follow the same strategy to define the sensitivity for the Analytic Gaussian mechanism and the approach proposed by [Brawner and Honaker \(2018\)](#). For the Wishart mechanism, the sensitivity is equal to the upper bound for l_2 -norm of the rows in $[X, Y]$. We therefore only consider item 2, i.e., an entry of S_H is set to be equal to zero if the same entry in S is zero by definition.

We also realize that the magnitude of the sensitivity depends on the upper and lower bounds of the variables (response and predictors) involved in the analysis. Since the variables are often in different scales, users can easily face scenarios where the magnitude of the sensitivity is highly dominated by, for example, a single variable. This single variable could have an interval length (upper minus lower bound) that is much larger than those of the remaining variables. If users ignore this issue under such scenario, the mechanism will add too much noise to those summaries involving the remaining variables. To avoid this problem, we first use the provided bounds to rescale all variables to lie in the $[0, 1]$ interval. We then implement the DP method and compute the point and interval estimates of the regression coefficients. Finally, we use the provided bounds again to scale back and report the estimates in the original scale.

11 Extended 2012 PUF Regression Results

For this section, we present additional findings from the regression experiments on the 2012 Public Use File (PUF).

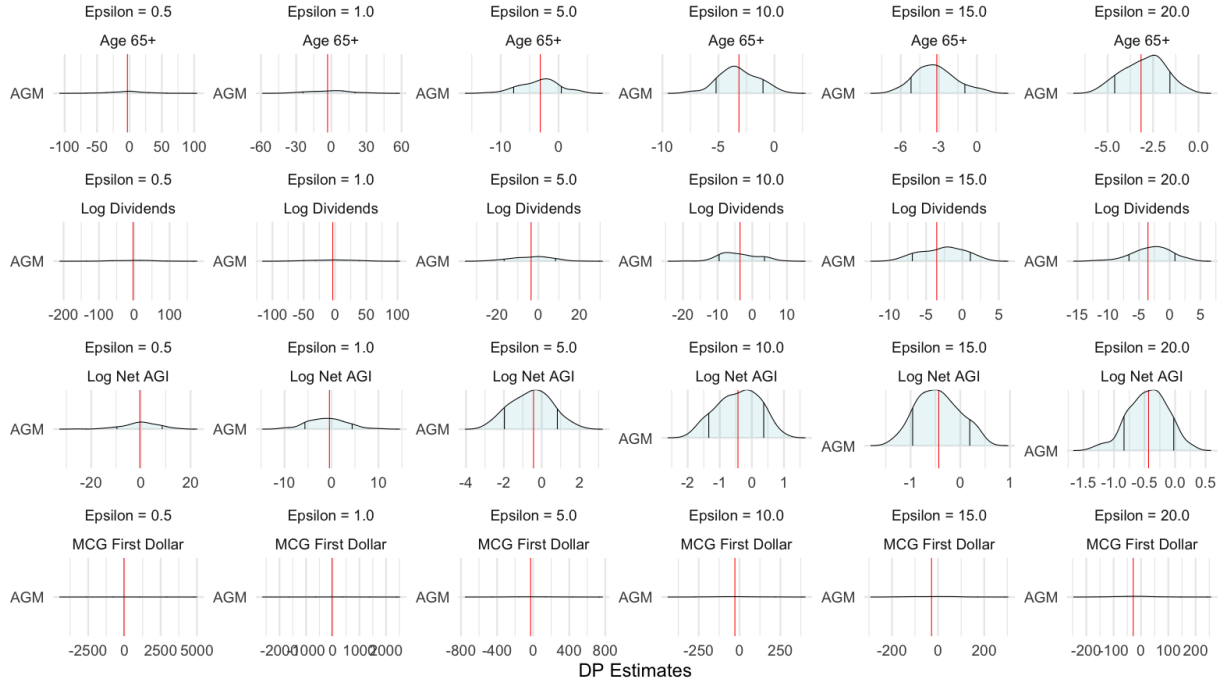


Figure 14: Distribution of absolute bias for regression coefficients produced by the Analytic Gaussian mechanism.

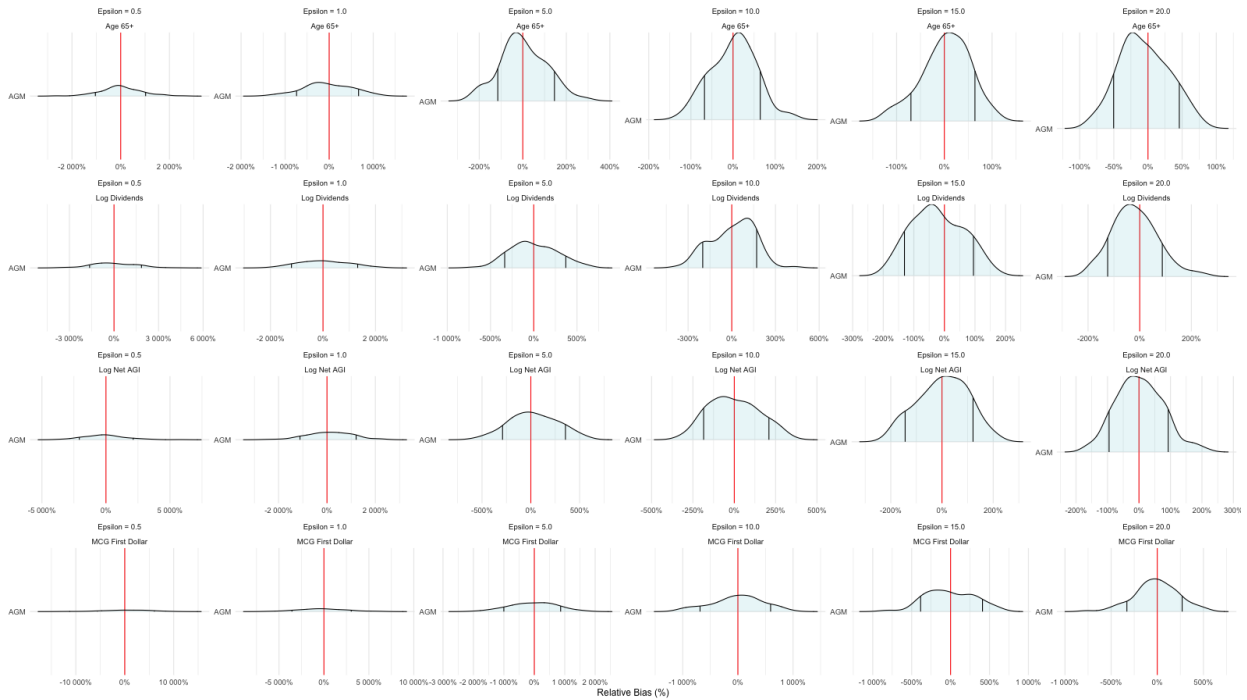


Figure 15: Distribution of relative bias for regression coefficients produced by the Analytic Gaussian mechanism.

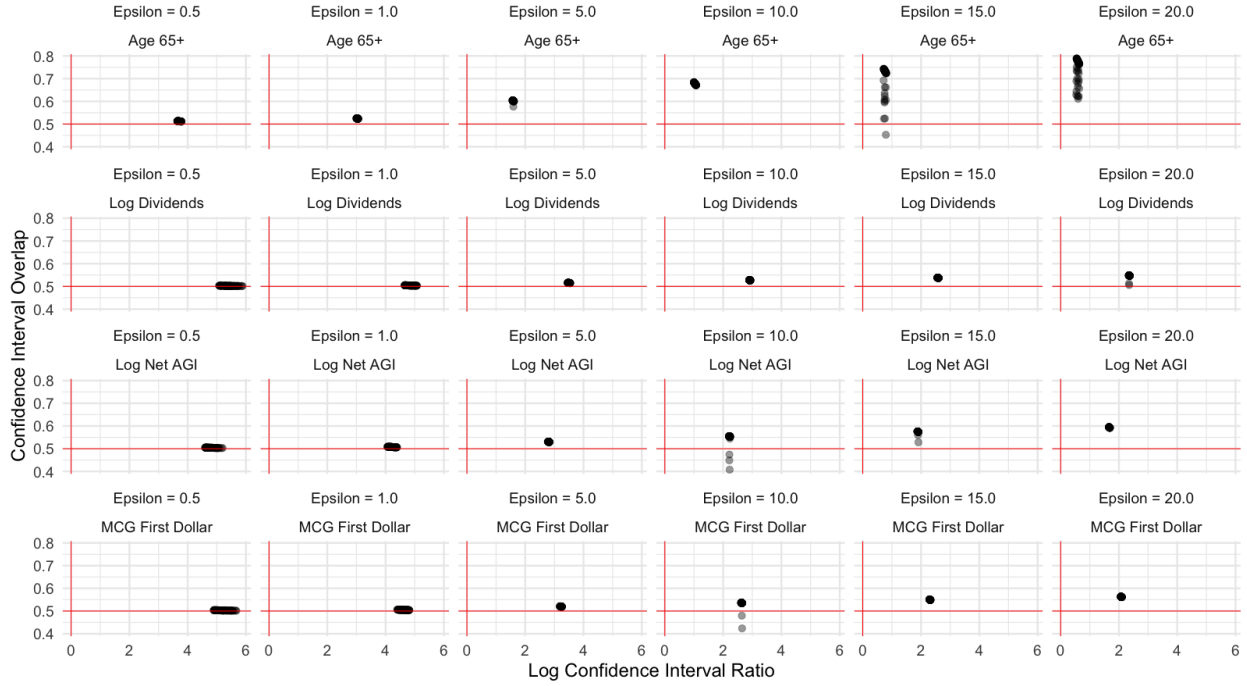


Figure 16: Confidence interval overlap vs. confidence interval ratios from the Analytic Gaussian mechanism for each regression coefficient using the bootstrap variability estimates. The vertical red lines indicate a CIR of 1 and the horizontal red lines indicate a CIO of 0.5.

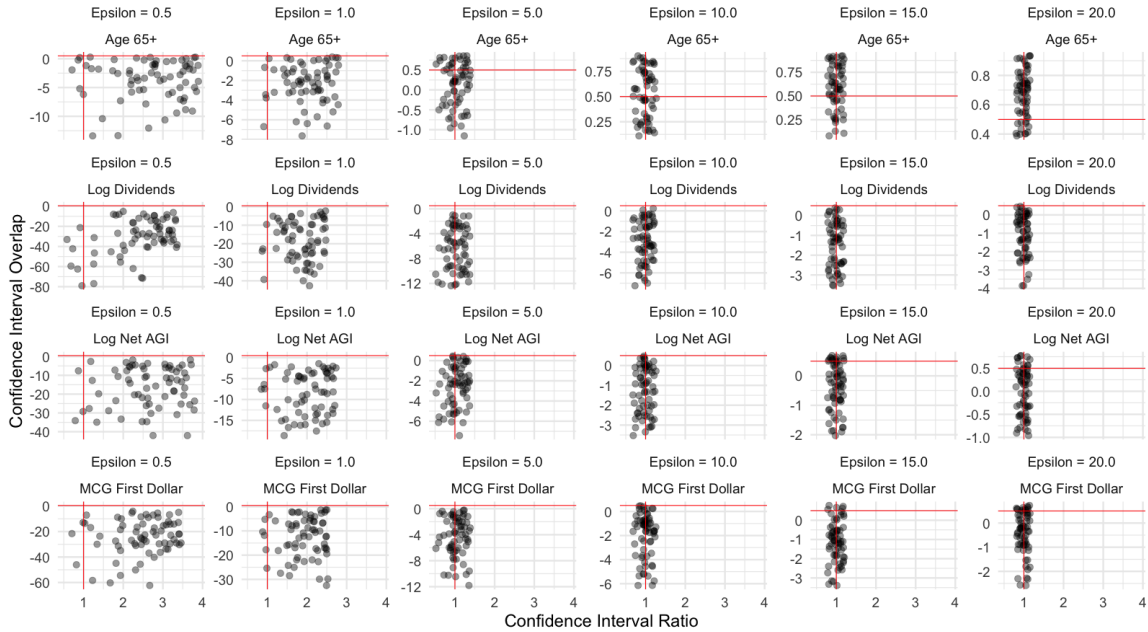


Figure 17: Confidence interval overlap vs. confidence interval ratios from the Analytic Gaussian mechanism for each regression coefficient using the asymptotic variability estimates. The vertical red lines indicate a CIR of 1 and the horizontal red lines indicate a CIO of 0.5.

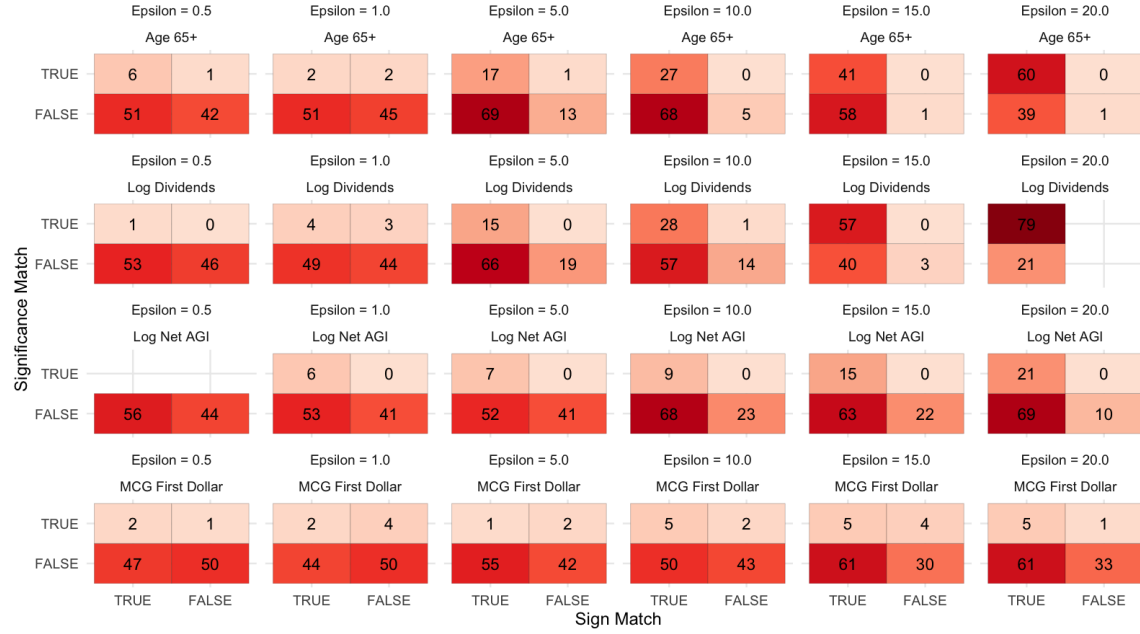


Figure 18: Confusion matrix of the Spherical Laplace mechanism results, showing percentages of sign and significance matches for each regression coefficient using the bootstrap confidence intervals. The upper left hand box indicates results that matched both.

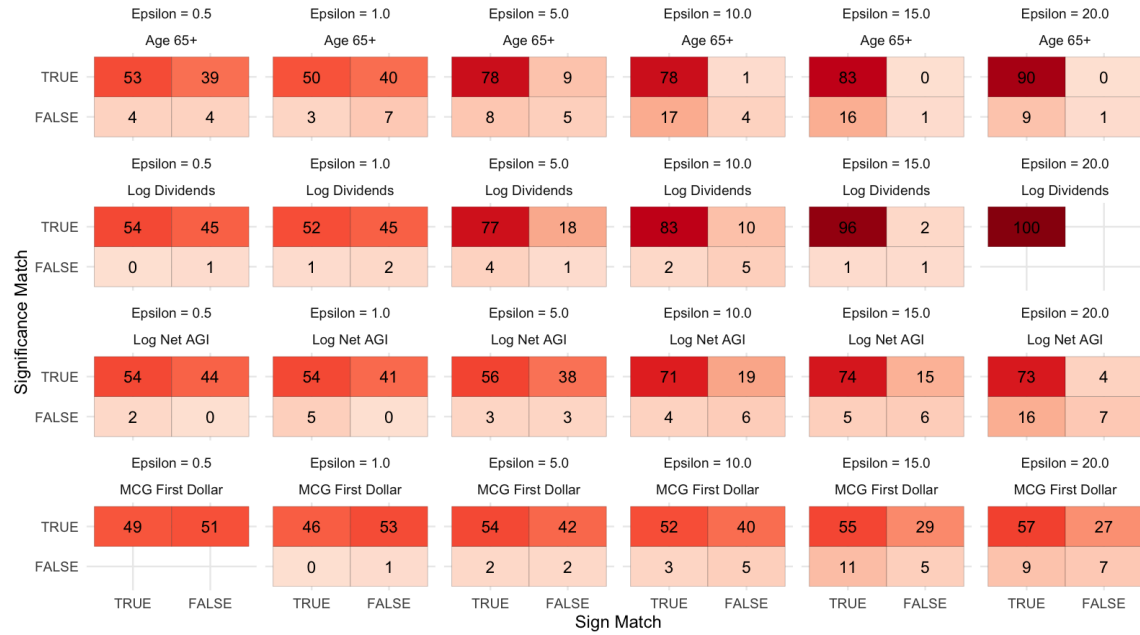


Figure 19: Confusion matrix of the Spherical Laplace mechanism results, showing percentages of sign and significance matches for each regression coefficient using the asymptotic confidence intervals. The upper left hand box indicates results that matched both.

12 CPS ASEC Results

We also tested our DP algorithms on the 1994 to 1996 Current Population Survey Annual Social and Economic Supplements (CPS ASEC) to ensure robustness and to have one dataset that is publicly accessible because Internal Revenue Service Statistics of Income (SOI) Division limits the access to the PUF. Additionally, CPS ASEC is one of the most similar microdata files to the PUF. The U.S. Census Bureau generates the CPS ASEC from a probability sample, which contains detailed information about income. Similar to the PUF, the CPS ASEC data has skewed variables and variables that are predominantly zeros. The one of the biggest differences between the PUF and CPS ASEC is that the latter reports information at the person level and household level instead of tax units. This difference should not have a meaningful impact on the feasibility study and will help test the flexibility of the DP methods. Moreover, unlike the PUF, the CPS ASEC represents the U.S. civilian non-institutional population and has 91,500 households and 157,959 people.

Note that the analysis from the [Mortenson and Whitten \(2020\)](#) does not work with the CPS ASEC data because the data are rounded and imprecise. However, we still calculate a histogram for demonstration using five years of pooled CPS ASEC data.

We also could not recreate the same analyses from [Feldstein et al. \(1980\)](#), because the CPS ASEC never contains dividends and only has capital gains as recently as 2008. There are complex ways around this but the researcher degrees of freedom do not seem warranted. We instead borrow a cross-sectional multiple linear regressions from [Card \(1999\)](#). These models are not causal and have been replaced by more sophisticated methods that generally report smaller effect sizes for the returns to education, but they are indicative of early quantitative publications on this topic. We reproduce Table 1 on page 1809 that aims to figure out what are the strong indicators of wage and salary income for gender.

13 CPS ASEC Regression Results

The complete CPS ASEC data and results can be found in the GitHub repo at [UrbanInstitute/formal-privacy-comp-appendix](#). Selected figures for the regression results are provided here for comparison with the SOI results reported in the main paper.

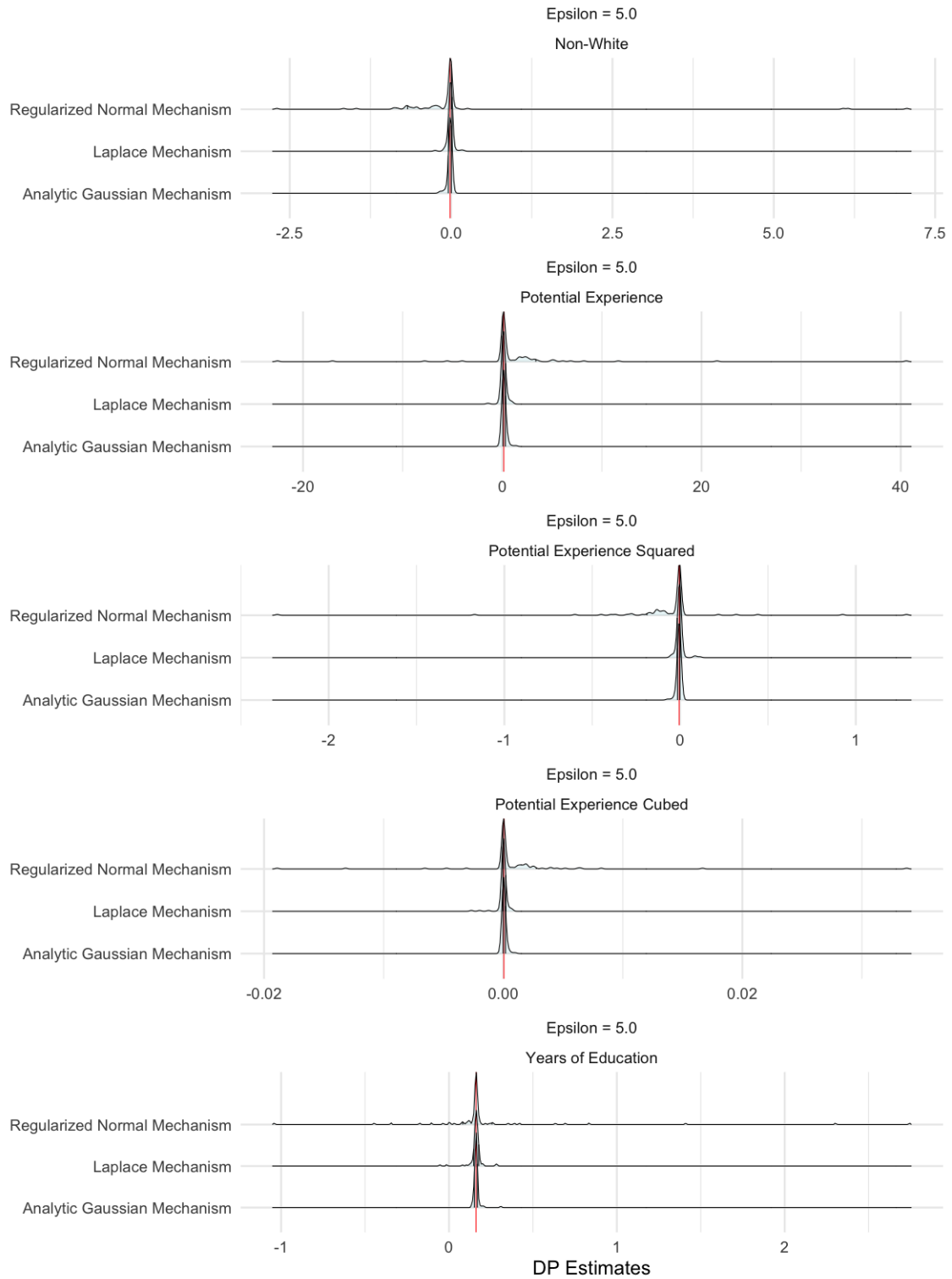


Figure 20: Distribution of Simulated DP Estimates for Regression Coefficients

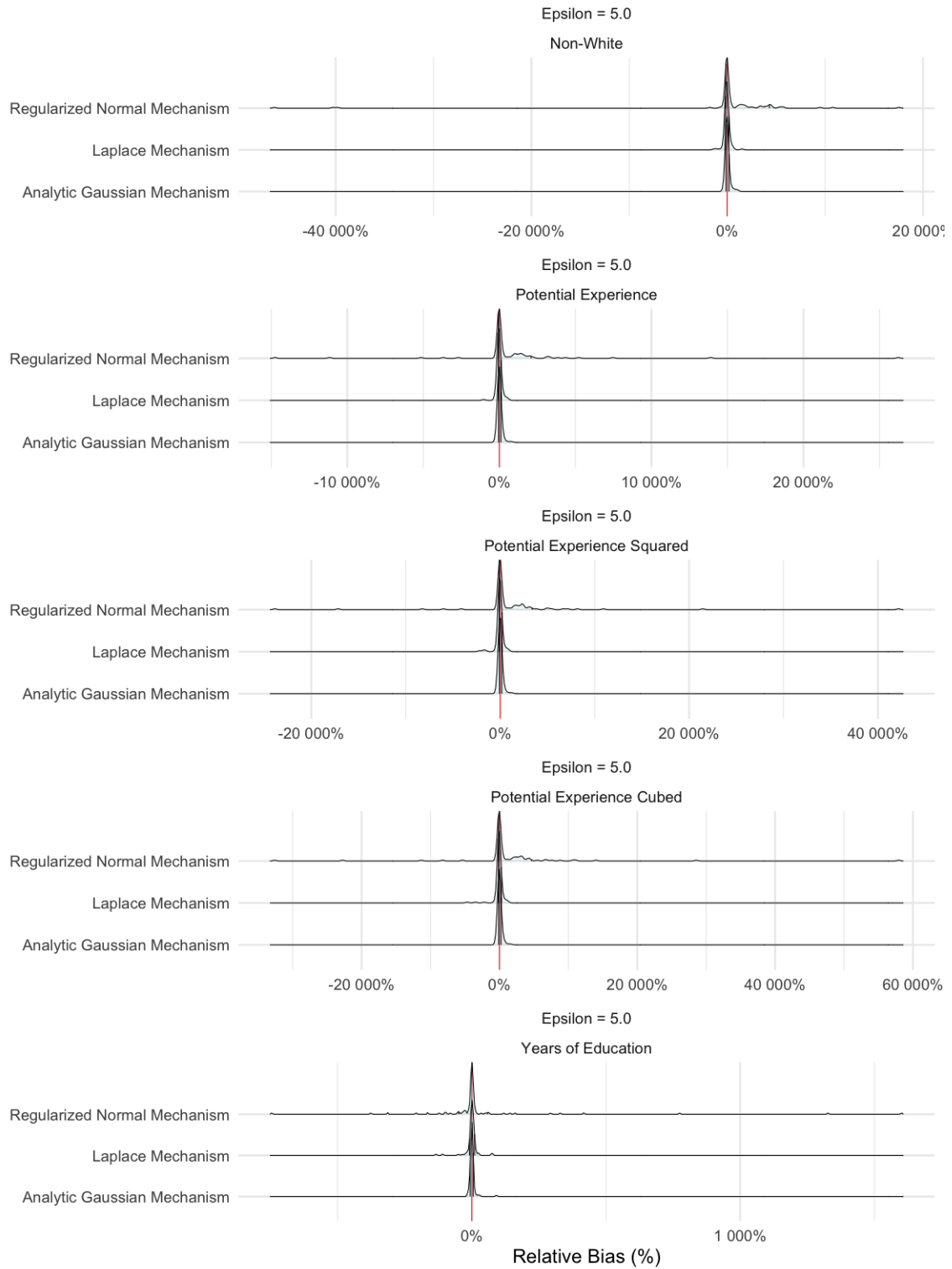


Figure 21: Distribution of Relative Bias for Regression Coefficients

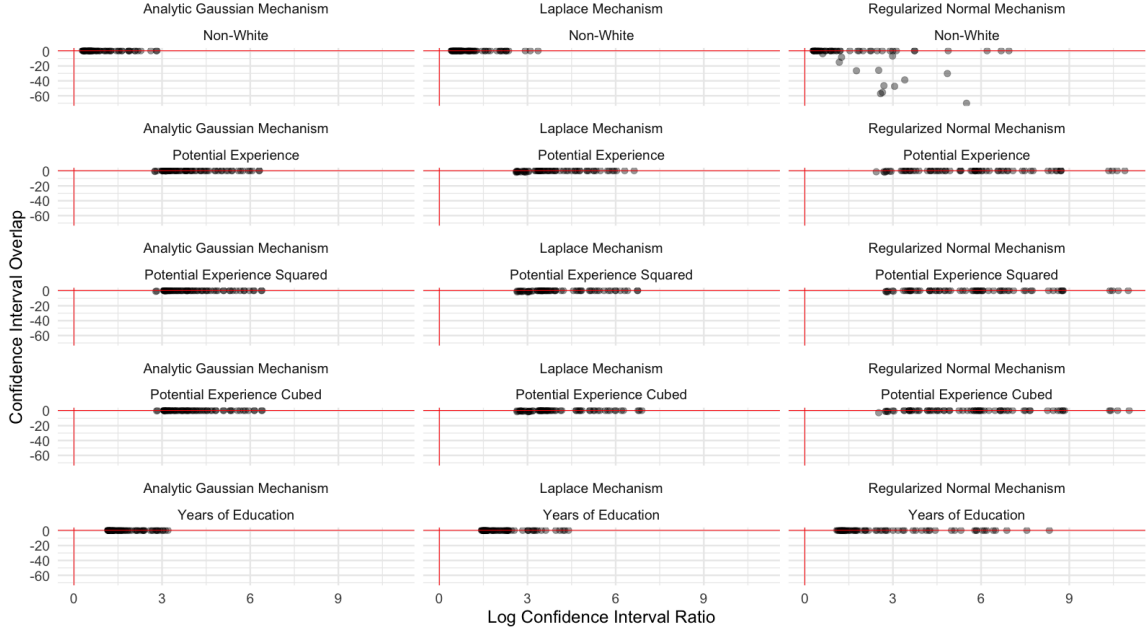


Figure 22: Confidence Interval Overlap vs. Confidence Interval Ratios for each regression coefficient using the bootstrap variability estimates. The vertical red lines indicate a CIR of 1 and the horizontal red lines indicate a CIO of 0.5. $\epsilon = 5$.

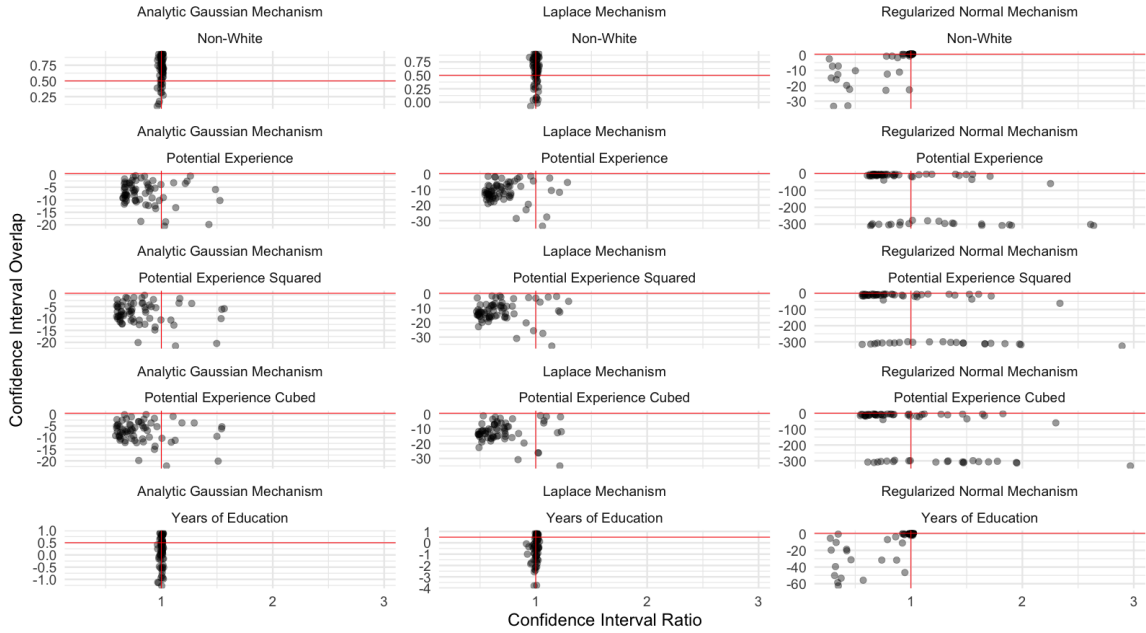


Figure 23: Confidence Interval Overlap vs. Confidence Interval Ratios for each regression coefficient using the asymptotic variability estimates. The vertical red lines indicate a CIR of 1 and the horizontal red lines indicate a CIO of 0.5. $\epsilon = 5$.

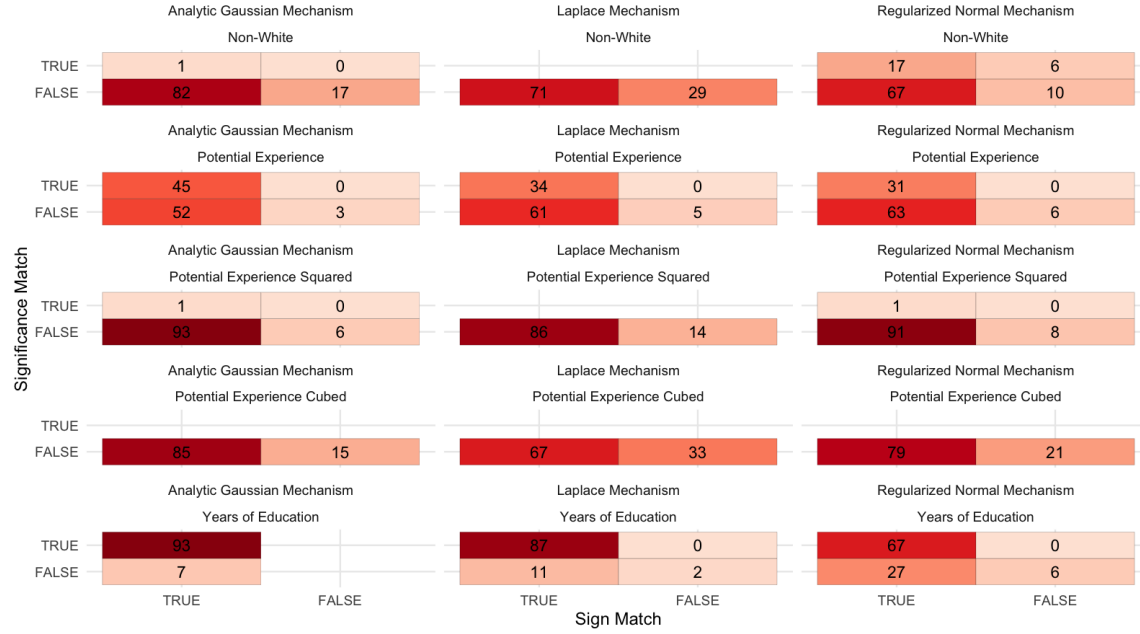


Figure 24: Confusion matrix showing percentages of sign and significance matches for each regression coefficient using the bootstrap confidence intervals. The upper left hand box indicates results that matched both. $\epsilon = 5$.

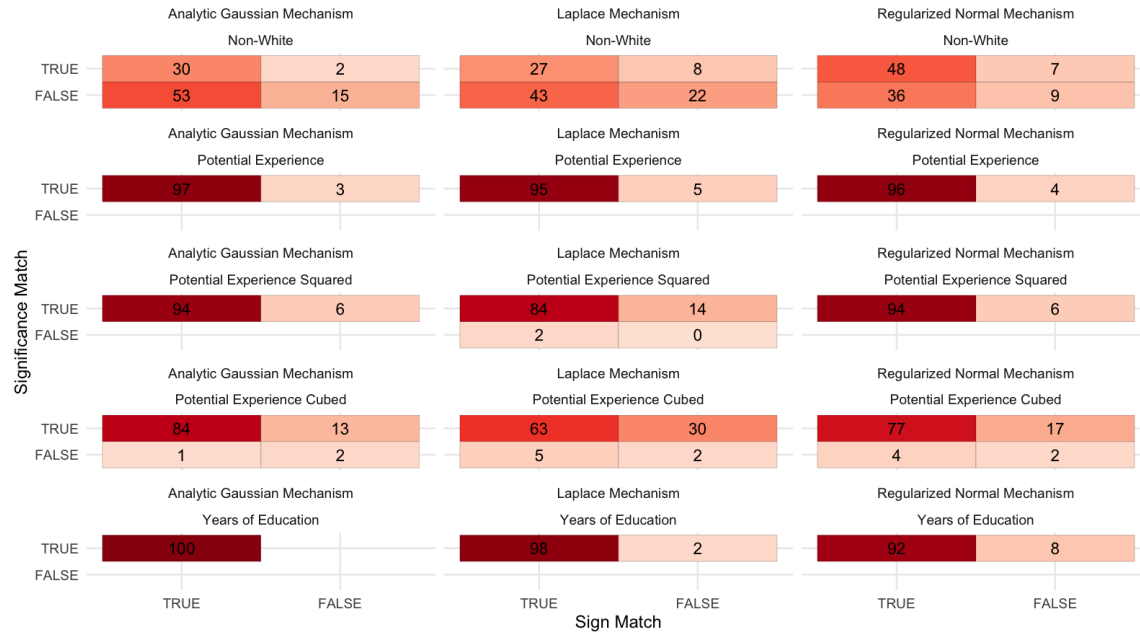


Figure 25: Confusion matrix showing percentages of sign and significance matches for each regression coefficient using the asymptotic confidence intervals. The upper left hand box indicates results that matched both. $\epsilon = 5$.

References

- Aktay, A., S. Bavadekar, G. Cossoul, J. Davis, D. Desfontaines, A. Fabrikant, E. Gabrilovich, K. Gadepalli, B. Gipson, M. Guevara, C. Kamath, M. Kansal, A. Lange, C. Mandayam, A. Oplinger, C. Pluntke, T. Roessler, A. Schlosberg, T. Shekel, S. Vispute, M. Vu, G. Wellenius, B. Williams, and R. J. Wilson (2020). Google covid-19 community mobility reports: Anonymization process description (version 1.0). *arXiv preprint arXiv:2004.04145*.
- Alabi, D., A. McMillan, J. Sarathy, A. Smith, and S. Vadhan (2020). Differentially private simple linear regression. *arXiv preprint arXiv:2007.05157*.
- Amitai, G. and J. Reiter (2018). Differentially private posterior summaries for linear regression coefficients. *Journal of Privacy and Confidentiality* 8(1).
- Avella-Medina, M. (2020). Privacy-preserving parametric inference: a case for robust statistics. *Journal of the American Statistical Association*, 1–15.
- 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. Bolton, T. Balmat, J. P. Reiter, J. M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, and M. DeLong (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the us federal government. *The Annals of Applied Statistics* 12(2), 1124–1156.
- Barrientos, A. F., J. P. Reiter, A. Machanavajjhala, and Y. Chen (2019). Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics* 28(2), 440–453.
- Benedetto, G., M. Stinson, and J. M. Abowd (2013). The Creation and Use of the SIPP Synthetic Beta.
- Bernstein, G. and D. R. Sheldon (2019). Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems* 32, 525–35.

- Biswas, S., Y. Dong, G. Kamath, and J. Ullman (2020). Coinpress: Practical private mean and covariance estimation. *Preprint, arXiv:2006.06618*.
- Bowen, C. M. and S. Garfinkel (2021). Philosophy of differential privacy. *Notices of the American Mathematical Society*.
- Bowen, C. M. and F. Liu (2020). Comparative study of differentially private data synthesis methods. *Statistical Science* 35(2), 280–307.
- Bowen, C. M., F. Liu, and B. Su (2021). Differentially private data release via statistical election to partition sequentially. *METRON* 79(1), 1–31.
- Brawner, T. and J. Honaker (2018). Bootstrap inference and differential privacy: Standard errors for free. unpublished manuscript.
- Bun, M. and T. Steinke (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–58. Springer.
- Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics* 3, 1801–1863.
- Chaudhuri, K., C. Monteleoni, and A. D. Sarwate (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12(3).
- Chen, E., Y. Miao, and Y. Tang (2020). Median regression with differential privacy. *arXiv preprint arXiv:2006.02983*.
- Chen, Y., A. Machanavajjhala, J. P. Reiter, and A. F. Barrientos (2016). Differentially private regression diagnostics. In *ICDM*, pp. 81–90.
- Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104(9), 2,633–79.
- Dandekar, A., D. Basu, and S. Bressan (2018). Differential privacy for regularised linear regression. In *International Conference on Database and Expert Systems Applications*, pp. 483–491. Springer.

- Dong, J., D. Durfee, and R. Rogers (2020). Optimal differential privacy composition for exponential mechanisms. In *International Conference on Machine Learning*, pp. 2,597–606. Proceedings of Machine Learning Research.
- D’Orazio, V., J. Honaker, and G. King (2015). Differential privacy for social science inference. (Available at SSRN 2676160).
- Drechsler, J. and L. Vilhuber (2014). Synthetic longitudinal business databases for international comparisons. In *International Conference on Privacy in Statistical Databases*, pp. 243–52. Springer.
- Du, W., C. Foot, M. Moniot, A. Bray, and A. Groce (2020). Differentially private confidence intervals. *Preprint, arXiv:2001.02285*.
- Dwork, C. (2008). Differential privacy: A survey of results. In *TAMC 2008, LNCS 4978*, pp. 1–19. Springer-Verlag Berlin Heidelberg.
- 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. and J. Lei (2009). Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380.
- 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*.
- Fang, X., F. Yu, G. Yang, and Y. Qu (2019). Regression analysis with differential privacy preserving. *IEEE access* 7, 129353–129361.

- Feldstein, M., J. Slemrod, and S. Yitzhaki (1980). The effects of taxation on the selling of corporate stock and the realization of capital gains. *The Quarterly Journal of Economics* 94(4), 777–791.
- 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*.
- Gong, M., K. Pan, and Y. Xie (2019). Differential privacy preservation in regression analysis based on relevance. *Knowledge-Based Systems* 173, 140–149.
- IBM (2019). Ibm differential privacy library.
- Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* 60(3), 224–232.
- Karwa, V. and S. Vadhan (2017). Finite sample differentially private confidence intervals. *Preprint, arXiv:1711.03908*.
- Lei, J., A.-S. Charest, A. Slavkovic, A. Smith, and S. Fienberg (2016). Differentially private model selection with penalized and constrained likelihood. *arXiv preprint arXiv:1607.04204*.
- 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.
- Liu, F. (2018). Generalized gaussian mechanism for differential privacy. *Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering* 31(4), 747–756.
- 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. and K. Talwar (2007). Mechanism design via differential privacy. In *48th Annual Institute of Electrical and Electronics Engineers Symposium on Foundations of Computer Science*, pp. 94–103. Institute of Electrical and Electronics Engineers.
- 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.
- Mortenson, J. A. and A. Whitten (2020). Bunching to maximize tax credits: Evidence from kinks in the us tax schedule. *American Economic Journal: Economic Policy* 12(3), 402–32.
- 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.
- Nissim, K., T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. R. O’Brien, and S. Vadhan (2017). Differential privacy: A primer for a nontechnical audience. In *Privacy Law Scholars Conference*.
- Peña, V. and A. F. Barrientos (2021). Differentially private methods for managing model uncertainty in linear regression models. *arXiv preprint arXiv:2109.03949*.
- Rinott, Y., C. M. O’Keefe, N. Shlomo, and C. Skinner (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science* 33(3), 358–385.
- Rogers, R., S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, and P. Ahammad (2020). LinkedIn’s audience engagements api: A privacy preserving data analytics system at scale. *arXiv preprint arXiv:2002.05839*.

- Ruggles, S., S. Flood, S. Foster, R. Goeken, J. Pacas, M. Schouweiler, and M. Sobek (2021). Ipums usa: Version 11.0 [dataset].
- 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.
- Shlomo, N. (2018). Statistical disclosure limitation: New directions and challenges. *Journal of Privacy and Confidentiality* 8(1).
- SmartNoise (2020). The exponential mechanism for medians.
- Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual Association for Computing Machinery Symposium on Theory of Computing*, pp. 813–22.
- Snoke, J. and C. M. Bowen (2019). Differential privacy: What is it? *AMSTAT News: The Membership Magazine of the American Statistical Association*, 26–8.
- Talwar, K., A. Guha Thakurta, and L. Zhang (2015). Nearly optimal private lasso. *Advances in Neural Information Processing Systems* 28, 3025–3033.
- Tax-Calculator (2021). Tax-calculator release, author’s calculation.
- Tzamos, C., E.-V. Vlatakis-Gkaragkounis, and I. Zadik (2020). Optimal private median estimation under minimal distributional assumptions. *arXiv preprint arXiv:2011.06202*.
- 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*.

- Wang, Y.-X., B. Balle, and S. P. Kasiviswanathan (2019). Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1,226–35. Proceedings of Machine Learning Research.
- Wasserman, L. and S. Zhou (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association* 105(489), 375–389.
- Zhang, J., Z. Zhang, X. Xiao, Y. Yang, and M. Winslett (2012). Functional mechanism: regression analysis under differential privacy. *arXiv preprint arXiv:1208.0219*.