Journal of Privacy and Confidentiality
Vol. 8 (1) 2018 Issue in honor of S. Fienberg

Submitted Sept 3, 2018
Published December 2018

# DIFFERENTIALLY PRIVATE POSTERIOR SUMMARIES FOR LINEAR REGRESSION COEFFICIENTS

GILAD AMITAI AND JEROME P. REITER

Department of Statistical Science, Duke University
*e-mail address*: gilad.amitai@duke.edu

Department of Statistical Science, Duke University
*e-mail address*: jreiter@duke.edu

Abstract. In Bayesian regression modeling, often analysts summarize inferences using posterior probabilities and quantiles, such as the posterior probability that a coefficient exceeds zero or the posterior median of that coefficient. However, with potentially unbounded outcomes and explanatory variables, regression inferences based on typical prior distributions can be sensitive to values of individual data points. Thus, releasing posterior summaries of regression coefficients can result in disclosure risks. In this article, we propose some differentially private algorithms for reporting posterior probabilities and posterior quantiles of linear regression coefficients. The algorithms use the general strategy of subsample and aggregate, a technique that requires randomly partitioning the data into disjoint subsets, estimating the regression within each subset, and combining results in ways that satisfy differential privacy. We illustrate the performance of some of the algorithms using repeated sampling studies. The non-private versions also can be used for Bayesian inference with big data in non-private settings.

## Introduction

Differential privacy (Dwork et al. [2006], Dwork and Roth [2014]) is now a gold standard criterion for protecting data subjects' privacy when releasing statistical outputs. Indeed, researchers have developed differentially private algorithms for many common statistical analyses, including counts and histograms, logistic regression, and various machine learning techniques. Many of these techniques focus on point estimation or prediction. Often, however, practitioners are interested in statistical inference for parameters of statistical models. Typically, these inferences involve confidence intervals (for frequentist inference) or summaries of posterior distributions (for Bayesian inference). In this article, we focus on statistical inference for coefficients in linear regression, where point estimates typically are not sufficient for interpretations of results.

*Key words and phrases:* differential privacy, Bayesian inference, posterior quantile, posterior probability.
This research was supported by the National Science Foundation grants SES-1131897 and ACI-1443014.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.683

> Steve Fienberg had an enormous influence on how I think about statistical science and a huge impact on my career. Steve's research is of course legendary; he made fundamental contributions to Bayesian inference, categorical data analysis, disclosure control, official statistics, and record linkage, to name key areas where our research interests overlapped. The statistical insights in this work are brilliant and important – and directly inspired several of my papers, including my first paper in *JASA*. The innovations in Steve's work target and respond to real, high impact problems in official statistics with leading edge methodology intended to make it into practice. His example of being an academic statistician who improves the practice of official statisticians on the ground is one that I try to emulate in my official statistics research.
>
> Although I was not one of Steve's students, I was the beneficiary of his mentorship. At conferences and workshops, and at reciprocating visits to CMU and Duke, Steve generously gave me many hours of his time, answering questions about research, official statistics, advising students, career options and paths, and life in general. I valued his advice dearly and followed it. I also have benefitted from Steve's mentorship in a more indirect way: I have been a post-doc mentor to two of his Ph.D. students. I can say from first hand experience that Steve's students are wonderful, both as scholars and as people. As I would have expected coming from Steve's tutelage.
>
> *Jerry Reiter*
>
> *DOI: 10.29012/jpc.692*

Relatively few algorithms exist for differentially private linear regression, and generally these do not enable statistical inference for coefficients; see Chen et al. [2018] for a review of differentially private algorithms for linear regression. In particular, most differentially private linear regression algorithms do not provide estimates of the standard errors of the coefficients. One notable exception is the method of Sheffet [2015], which can be used for interval estimation. However, this algorithm actually reports intervals from a ridge regression with positive probability rather than a linear regression always. The algorithm assumes known bounds on the outcome and all explanatory variables, and appears to require $(\epsilon, \delta > 0)$-differential privacy rather than just $\epsilon$-differential privacy.

In this article, we propose approaches for statistical inference for linear regression coefficients. In particular, we provide $\epsilon$-differentially private summaries of Bayesian posterior probabilities and Bayesian posterior quantiles, without assuming bounds on the outcome and explanatory variables. The algorithms are based on the general strategy of subsample and aggregate presented in Nissim et al. [2007]. We first randomly split the private dataset into disjoint partitions. In each partition, we estimate the linear regression of interest. We combine quantities from each individual regression, adding noise using Laplace mechanisms to ensure differential privacy.

The remainder of this article is organized as follows. In Section 1 we briefly review differential privacy. In Section 2 we describe two algorithms for releasing differentially private posterior probabilities. In Section 3, we evaluate the empirical performance of the algorithms for releasing differentially private posterior probabilities. In Section 4, we outline a method for releasing differentially private posterior quantiles. We do not evaluate the performance of this algorithm here; this will be the subject of future work. Finally, in Section 5, we conclude with overall findings and suggested topics for further study.

## 1. Review of Differential Privacy

Let $\mathcal{A}$ be an algorithm that takes as input a database $\mathcal{D}$ and outputs some quantity $o$, i.e., $\mathcal{A}(\mathcal{D}) = o$. In our context, these outputs include summaries of the posterior distributions of linear regression coefficients. Differential privacy uses the concept of neighboring databases.

For this article, as done for most differentially private algorithms in the literature, we define neighboring databases $\mathcal{D}$ and $\mathcal{D}^*$ as different in one row and identical for all other rows.

**Definition 1.1.** [$\epsilon$-differential privacy] An algorithm $\mathcal{A}$ satisfies $\epsilon$-differential privacy, abbreviated $\epsilon$-DP, if for any pair of neighboring databases $(\mathcal{D}, \mathcal{D}^*)$, and any set $S \in range(\mathcal{A})$, $Pr(\mathcal{A}(\mathcal{D}) \in S) \leq \exp(\epsilon) Pr(\mathcal{A}(\mathcal{D}^*) \in S)$.

A common method for ensuring $\epsilon$-DP is the Laplace mechanism. For any function $f : \mathcal{D} \to \mathbb{R}^d$, let $\Delta(f) = \max_{(\mathcal{D},\mathcal{D}^*)} ||f(\mathcal{D}) - f(\mathcal{D}^*)||_1$, where $(\mathcal{D}, \mathcal{D}^*)$ are neighboring databases. For example, for counting the number of people with a certain property in a database, $f$ represents summing a binary indicator variable with $\Delta(f) = 1$. For any $f$ with global sensitivity $\Delta(f)$, the Laplace mechanism is

$$\mathbf{LM}(\mathcal{D}) = f(\mathcal{D}) + \eta, \tag{1.1}$$

where $\eta$ is a $d \times 1$ vector of independent draws from a Laplace distribution with density $p(x \mid \lambda) = (1/(2\lambda)) \exp(-|x|/\lambda)$, where $\lambda = \Delta(f)/\epsilon$.

As a graduate student, I was nervous when giving my first research presentation to the US Census research group at Carnegie Mellon's Department of Statistics: Steve Fienberg was in the room, and I was not sure if my record linkage research was any good. While I was presenting, Steve seemingly paid no attention to what I was saying. I would glance his direction when presenting each slide, and every time, he was typing something on his iPad and not watching my presentation at all. That's okay, I thought. He's probably really busy, and shouldn't be wasting his time on my talk.

After the meeting, I got back to my desk and checked my inbox, only to find seven emails from Steve, each with a paper that I should read about record linkage that related to the different aspects of my research. His feedback in these emails helped shape the directions of my future research.

That was Steve. He always made time for sharing constructive feedback on student research, he always knew more than you thought he did about your research, and he could multitask better than anyone I've ever met.

*Sam Ventura*
*DOI: 10.29012/jpc.693*

For some $f$, $\Delta(f)$ can be large compared to $f(\mathcal{D})$. As a result, the Laplace mechanism may add so much noise that the realized value of $\mathbf{LM}(\mathcal{D})$ is too far from $f(\mathcal{D})$ to be useful. For example, when $f(\mathcal{D})$ is an estimated coefficient in a linear regression with (approximately) continuous-valued outcomes or explanatory variables, high leverage points and outliers have the potential to change the values of regression coefficients by huge amounts, resulting in impractical values of $\Delta(f)$. Similarly, when $f(\mathcal{D})$ is a probability summarizing the distribution of an estimated regression coefficient, individual data points could swing the probability from near one to near zero, or vice versa. With $\Delta(f) = 1$, the Laplace mechanism is likely to completely destroy the usefulness of the reported probability.

In such cases, it can be useful to control the global sensitivity of $f$ by using the subsample and aggregate technique (Nissim et al. [2007]). We partition $\mathcal{D}$ into $M$ exclusive subsets, $\mathcal{D}_1, \dots, \mathcal{D}_M$. In each $\mathcal{D}_j$, where $j = 1, \dots, M$, we compute some function $h(\mathcal{D}_j)$. We then compute some aggregation function $g(h(\mathcal{D}_1), \dots, h(\mathcal{D}_M))$ that can be used to approximate $f(\mathcal{D})$. As a simple illustration, when $f(\mathcal{D}) = \sum_{i=1}^n y_i/n$, where $n$ is the sample size of $\mathcal{D}$, we could set $h(\mathcal{D}_j) = \sum_{i \in \mathcal{D}_j} y_i/(n/M)$ and $g(h(\mathcal{D}_1), \dots, h(\mathcal{D}_M)) = \sum_{j=1}^M h(\mathcal{D}_j)/M = f(\mathcal{D})$. Analysts typically choose $g(h(\mathcal{D}_1), \dots, h(\mathcal{D}_M))$ so that its global sensitivity is less than $\Delta(f)$. In particular, any one data point appears in only one $\mathcal{D}_j$. As a result, using subsample and aggregate with $g$ can allow the analyst to reduce global sensitivity by a factor of $1/M$ compared to the global sensitivity of $f$. Once $\Delta(g)$ is determined, analysts can use the Laplace mechanism to perturb $g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M))$. The output of the noisy aggregation function can go through post-processing steps to approximate $f(\mathcal{D})$.

## 2. Differentially Private Posterior Probabilities

Let the confidential dataset $\mathcal{D}$ comprise $n$ individuals, $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$, where the analyst treats $y_i$ as the dependent variable and $\boldsymbol{x}_i$ as the $(p+1) \times 1$ vector of explanatory variables for individual $i$. The first element of each $\boldsymbol{x}_i$ equals one for all $i$, which encodes the intercept term in a linear regression. We seek to estimate the model,

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \omega_i \quad \omega_i \sim \mathcal{N}\left(0, \sigma^2\right), \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ is a vector of regression coefficients and $\sigma^2$ is the regression variance.

We estimate the parameters in (2.1) using Bayesian methods. In this article, we use the popular Zellner's $g$ prior distribution [Hoff, 2009, p. 157] for $(\boldsymbol{\beta}, \sigma^2)$, which facilitates computation. Conditional on $\sigma^2$, the resulting posterior distribution for $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\sigma^2, \mathcal{D}) \sim \mathcal{N}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}, \frac{g}{g+1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2\right). \tag{2.2}$$

Here, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{y})$, where $\boldsymbol{X}$ is the $n \times (p+1)$ matrix of explanatory variables and $\boldsymbol{y}$ the $n \times 1$ vector of responses. The constant $g$ often is selected to equal the sample size $n$, so that for large $n$ it is practically irrelevant. We average the distribution in (2.2) over the posterior distribution of $\sigma^2$, which is an inverse-gamma distribution, to get $p(\boldsymbol{\beta}|\mathcal{D})$. For large $n$, it is often reasonable to fix $\sigma^2$ at its posterior mode $\hat{\sigma}^2$, in which case we simply use (2.2) with $\hat{\sigma}^2$ plugged in for $\sigma^2$. We note that our results apply for arbitrary prior distributions, provided that the information in the data swamps the information in the prior distribution.

For any particular $\beta_k \in \boldsymbol{\beta}$, we write the marginal posterior distribution (conditional on $\sigma^2$) as

$$p(\beta_k|\sigma^2, \mathcal{D}) \sim \mathcal{N}\left(\hat{\beta}_k, v_k\right). \tag{2.3}$$

Here, $\hat{\beta}_k$ is the $k$th entry of the posterior mode of $\boldsymbol{\beta}$, and $v_k$ is the $k$th diagonal entry of $\mathrm{Var}\left[\boldsymbol{\beta}|\sigma^2, \mathcal{D}\right]$.

Bayesian inferences for $\beta_k$ often focus on posterior probabilities, $p(b) = \Pr\left(\beta_k \leq b|\mathcal{D}\right)$, where $b$ is some analyst-specified constant. For example, $p(0)$ represents the probability that $\beta_k$ is negative. Since the global sensitivity of $p(b)$ equals one, a direct application of the Laplace mechanism adds so much noise as to totally obfuscate the information in $p(b)$. We therefore approximate $p(b)$ using two algorithms based on the subsample and aggregate technique. The approach in Section 2.1, which we call the *rescaled-normal* approach, uses special features of the normal distribution to approximate $p(b)$. The approach in Section 2.2, which we call the *Fisher-derived* approach, is based on Fisher's method for combining $p$-values (Fisher [1925]) that we modify for approximating $p(b)$.

2.1. **Rescaled-Normal Approach.** Given only a partition $\mathcal{D}_j = \{\boldsymbol{y}_j, \boldsymbol{X}_j\}$, and using the Zellner $g$ prior distribution determined using only quantities from $\mathcal{D}_j$, the analyst would compute

$$p(\boldsymbol{\beta}|\sigma^2, \mathcal{D}_j) \sim \mathcal{N}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, (\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\sigma^2\right). \tag{2.4}$$

Here, $\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}(\boldsymbol{X}_j^T\boldsymbol{y}_j)$, where $\boldsymbol{X}_j$ is the $n_j \times (p+1)$ matrix of explanatory variables and $\boldsymbol{y}_j$ the $n \times 1$ vector of responses from $\mathcal{D}_j$. We again assume $g$ is large enough so as to be practically irrelevant for posterior computations. We write the marginal posterior distribution of any $\beta_k$ given $\mathcal{D}_j$ as

$$p(\beta_k|\sigma^2, \mathcal{D}_j) \sim \mathcal{N}\left(\hat{\beta}_{j,k}, v_{j,k}\right), \tag{2.5}$$

where $\hat{\beta}_{j,k}$ is the $k$th entry of $\hat{\boldsymbol{\beta}}_j$ and $v_{j,k}$ is the $k$th diagonal entry of Var $\left[\boldsymbol{\beta}|\sigma^2, \mathcal{D}_j\right]$. From here it is straightforward to compute any $p_j(b) = p(\beta_k \leq b|\mathcal{D}_j)$.

After partitioning $\mathcal{D}$, we seek to aggregate the $M$ values of $p_j(b)$. Simply averaging all $M$ values, i.e., using $\bar{p}(b) = \sum_{j=1}^{M} p_j(b)/M$, does not result in a reliable estimate of $p(b)$, even absent noise addition. This is because the posterior variance in (2.5) tends to exceed the posterior variance in (2.3), as the former is based on a smaller sample size than the latter. The higher variability tends to result in larger probability mass in tail areas, making $\bar{p}(b)$ generally dissimilar to $p(b)$.

To avoid this inconsistency, we seek to adjust $\bar{p}(b)$. To do so, we first assume that we are working with a large enough $n$ so that (i) the posterior variance of $\boldsymbol{\beta}$ using the full $\mathcal{D}$ can be approximated with $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$, and (ii) the posterior variance of $\boldsymbol{\beta}$ using any $\mathcal{D}_j$ can be approximated with $(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\sigma^2$. In other words, the data swamp the effects of the prior distribution, and it is reasonable to treat $\sigma^2$ as known. Given approximate normality of the posterior distribution, $p_j(b) = \Phi\left((b - \hat{\beta}_{j,k})/\sqrt{v_{j,k}}\right)$, where $\Phi()$ indicates the cumulative probability associated with the argument under the standard normal distribution.

We assume that the differences among the posterior variances for each partition are small enough that we can approximate all of them accurately with a common variance $\tilde{v}_k$. Making this assumption, we have

$$\bar{p}(b) = \frac{1}{M}\sum_{j=1}^{M}\Phi\left(\frac{b - \hat{\beta}_{j,k}}{\sqrt{v_{j,k}}}\right) \approx \frac{1}{M}\sum_{j=1}^{M}\Phi\left(\frac{b - \hat{\beta}_{j,k}}{\sqrt{\tilde{v}_k}}\right). \tag{2.6}$$

Posterior variances often have lower variability than posterior means ([Rubin, 1987, Ch. 3]), particularly in large samples, which motivates why it can be reasonable to assume constant posterior variances across $\mathcal{D}_j$.

It is difficult to manipulate the sums in (2.6), because each normal probability has a different mean. To get around this difficulty, we use the approximation,

$$\frac{1}{M}\sum_{j=1}^{M}\Phi\left(\frac{b - \hat{\beta}_{j,k}}{\sqrt{\tilde{v}_k}}\right) \approx \Phi\left(\frac{b - \hat{\beta}_k}{\sqrt{\tilde{v}_k}}\right). \tag{2.7}$$

The rationale of this approximation is as follows. When averaging over repeated samples of $\mathcal{D}$ and the partitioning process, the $E(\hat{\beta}_{j,k}) = E(\hat{\beta}_k) = \beta$. Hence, for large samples, it may not sacrifice accuracy too much to assume that the individual $p_j(b)$'s tend to be centered at a posterior probability computed with $\hat{\beta}_k$. Of course, $\Phi()$ is not a linear function of $\hat{\beta}_k$ nor

$\hat{\beta}_{j,k}$, so that the approximation can be inaccurate particularly for small samples. We leave further analytical study of the validity of this approximation to future work.

Finally, we rescale the variance in the last term of (2.7) to match the variance of $\hat{\beta}_k$. To do so, we assume that $v_k \approx \tilde{v}_k/M$. In other words, since $\mathcal{D}_j$ includes a random sample of $(1/M) \times 100$ percent of the individuals in $\mathcal{D}$, it is reasonable in large samples to approximate $\tilde{v}_k$ as $M$ times larger than the value of $v_k$. Mathematically, as sample sizes increase, when $\mathcal{D}_j$ is a randomly sampled partition, $M\boldsymbol{X}_j^T\boldsymbol{X}_j$ should approach $\boldsymbol{X}^T\boldsymbol{X}$. With this approximation, we have

$$\sqrt{M}\Phi^{-1}(\bar{p}(b)) \approx \sqrt{M}\left(\frac{b-\hat{\beta}_k}{\sqrt{\tilde{v}_k}}\right) \approx \frac{b-\hat{\beta}_k}{\sqrt{v_k}}. \tag{2.8}$$

Thus, we arrive at an expression with a variance appropriate for a computation based on $\mathcal{D}$, as desired. We compute the cumulative probability associated with the statistic in (2.8) as an approximation of $p(b)$.

Turning now to making this approximation differentially private, we use the subsample and aggregate technique. Each $h(\mathcal{D}_j) = p_j(b)$, and

$$g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) = \frac{1}{M}\sum_{j=1}^{M} p_j(b). \tag{2.9}$$

The global sensitivity of $g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M))$ is $1/M$. We use the Laplace mechanism to perturb $g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M))$, and use post-processing to compute the differentially private approximation to $p(b)$, which we call $p_{N,\epsilon}(b)$, where the subscript indicates the rescaled-normal method with privacy budget $\epsilon$. Putting it all together, we have

$$p_{N,\epsilon}(b) = \Phi\left(\sqrt{M}\cdot\Phi^{-1}\left(\sum_{j=1}^{M} p_j(b)/M + \eta\right)\right), \tag{2.10}$$

where $\eta$ is a draw from the Laplace distribution with mean parameter zero and scale parameter $\epsilon/M$ with $\epsilon$ equal to the privacy budget. The final algorithm is displayed below.

**Rescaled-normal Algorithm.** *Inputs: $\mathcal{D}$, $M$, $\epsilon$, Model, $\beta_k$, $b$. Output: $p_{N,\epsilon}(b)$.*

(1) *Partition $\mathcal{D}$ into $M$ disjoint subsets, $(\mathcal{D}_1, \ldots, \mathcal{D}_M)$.*
(2) *In each $\mathcal{D}_j$, where $j = 1, \ldots, M$, fit the Bayesian regression specified in the input Model, and estimate the posterior probability $p_j(b) = \Pr(\beta_k \leq b|\mathcal{D}_j)$.*
(3) *Draw $\eta$ from a Laplace distribution with scale parameter $\epsilon/M$.*
(4) *Compute $p_{N,\epsilon}(b) = \Phi\left(\sqrt{M}\cdot\Phi^{-1}\left(\sum_{j=1}^{M} p_j(b)/M + \eta\right)\right)$,*

We close this section with a summary of the key assumptions and conditions underlying the rescaled-normal approach. The most critical is that the posterior distribution of $\beta_k$ is approximately a normal distribution. When this is not reasonable, e.g., when sample sizes are small and the input model is a poor description of the data, one should not count on $p_{N,\epsilon}(b)$ to provide meaningful results. In general, it is difficult to check the validity of model assumptions without access to individual data values, which generally are private and hence not available to the data analyst. This difficulty affects any differentially private regression algorithm. One partial solution is to preserve some privacy budget to allow examination of differentially private residual plots to assess the validity of the regression model assumptions [Chen et al., 2018].

We make several approximations about the set of $p_j(b)$ values to get to (2.8). To get to (2.6), we assume that the posterior variances of $\beta_k$ from $M$ randomly sampled datasets are similar enough that practically one can call them equal. We expect the performance of the algorithm to be insensitive to minor violations of this assumption—which is what one would expect in large samples—as the posterior probabilities do not change substantially for small changes to the posterior variances. To get to (2.7), we assume that the average of the posterior probabilities from $M$ random samples has expectation equal to the posterior probability computed with $\hat{\beta}_k$. We expect the algorithm to be sensitive to this assumption when variability in the values of $\hat{\beta}_{j,k}$ is large. This results in large variability in the values of $p_j(b)$, which may cause the average in (2.7) to pull away from the value on the right hand side of (2.7). Finally, to get to (2.8), we assume that the posterior variance in a dataset of size $n/M$, i.e., the partitions, is approximately $M$ times larger than the posterior variance in a dataset of size $n$. We expect the algorithm to be quite sensitive to this assumption, as it is crucial for translating the results from the partitions to the results from the full dataset. This assumption should be reasonable in large samples, provided that there are not individual data points with extremely high leverage compared to the rest of the observations.

2.2. **Fisher-derived Method.** Combining probabilities across independent datasets is reminiscent of Fisher's method (Fisher [1925]) for combining $p$-values from independent significance tests. Suppose that one does $M$ independent tests of a common null hypothesis, for example, $H_0 : \beta_k = 0$. In Fisher's method, these tests come from $M$ disjoint datasets, such as replications of a study done at $M$ locations. The goal of Fisher's method is to combine the $M$ $p$-values into one omnibus test of $H_0$, which is derived as follows. When $H_0$ is true, $p$-values follow a uniform distribution on $(0, 1)$. Further, for any uniformly distributed random variable $U$, $-2\log(U)$ has a chi-squared distribution with 2 degrees of freedom, denoted $\chi^2_2$. Hence, for a sum of $M$ independent uniformly distributed random variables $(U_1, \ldots, U_M)$, using the properties of sums of chi-squared random variables we have

$$-2\sum_{j=1}^{M}\log(U_j) \sim \chi^2_{2M}. \tag{2.11}$$

Fisher's test uses the $M$ $p$-values in the sum in (2.11) as the test statistic. The area under the $\chi^2_{2M}$ distribution to the right of the statistic is the $p$-value for the omnibus test of $H_0$.

Although $p$-values and posterior probabilities derive from entirely different philosophies, they can be closely connected for many scalar estimands. In particular, suppose that one assumes the null hypothesis $H_0 : \beta_k = b$ and a one-sided alternative hypothesis, say $H_0 : \beta_k > b$. Then, for large samples, the frequentist $p$-value of the test of $H_0$ equals $1 - \Phi((\hat{\beta}_k - b)/\sqrt{v_k}) = \Phi((b - \hat{\beta}_k)/\sqrt{v_k})$. With diffuse prior distributions, this is exactly the same value mathematically as $p(b)$.

Given this connection, we propose to adapt Fisher's method for combining posterior probabilities. In general, posterior probabilities do not follow uniform distributions (except when $\beta = b$); thus, this adaptation of Fisher's method for Bayesian posterior summaries is a heuristic. However, we can evaluate the properties of this approach in finite samples via simulation, as we do in Section 3.

To define the method more formally, for each partition $j$ let $h(\mathcal{D}_j) = -2\log(p_j(b))$. We are working with the logarithms of probabilities, which are bounded between zero and negative infinity; thus, $h(\mathcal{D}_j)$ has infinite global sensitivity. To bound the sensitivity, we clip each $-2\log(p_j(b))$ at an analyst-specified value. The analyst defines a probability threshold $p^*$ such that the analyst is comfortable with treating posterior probabilities smaller than $p^*$ as equivalent to $p^*$; for example, $p^* = .001$ might suffice. As a consequence, the analyst cannot distinguish posterior probabilities below $p^*$. The global sensitivity of any $-2\log(p_j(b))$ is then $-2\log(p^*)$.

To make the differentially private test statistic, we let $\tilde{p}_j(b) = p_j(b)$ when $-2\log(p_j(b)) \leq -2\log(p^*)$, and $\tilde{p}_j(b) = p^*$ otherwise. Let

$$g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) = -2\sum_{j=1}^{M}\log(\tilde{p}_j(b)). \tag{2.12}$$

This has global sensitivity $-2\log(p^*)/M$. Using the Laplace mechanism, we compute a differentially private version of $g$ using

$$F_\epsilon = -2\sum_{j=1}^{M}\log(\tilde{p}_j(b)) + \eta \tag{2.13}$$

where $\eta$ is a draw from the Laplace distribution with mean parameter zero and scale parameter $-2\log(p^*)/(M\epsilon)$.

The reference distribution for $F_\epsilon$ is not available in analytical form. However, under the null hypothesis $H_0 : \beta = b$, we can generate a reference distribution that takes into account the clipped nature of the posterior probabilities and the addition of noise as follows.

(1) Sample $M$ posterior probabilities, say $u_1, \ldots, u_M$, from a standard uniform distribution.
(2) For $j = 1, \ldots, M$, compute $\tilde{p}(u_j) = -2\log(u_j)$ when $-2\log(u_j) \leq -2\log(p^*)$ and $\tilde{p}(u_j) = -2\log(p^*)$ when $-2\log(u_j) > -2\log(p^*)$.
(3) Sample a draw $\eta$ from a Laplace distribution with scale parameter $-2\log(p^*)/(M\epsilon)$.
(4) Compute $-2\sum_{j=1}^{M}\log(\tilde{p}(u_j)) + \eta$.
(5) Repeat steps 1–4 thousands of times. The draws approximate the reference distribution of $F_\epsilon$. We recommend using at least 10000 draws.

The estimate of $p(b)$, which we call $p_{F,\epsilon}(b)$, is determined as the proportion of the Monte Carlo draws from the reference distribution that exceed $F_\epsilon$. We summarize the Fisher-derived algorithm below.

**Fisher-derived Algorithm.** *Inputs: $\mathcal{D}$, $M$, $\epsilon$, $p^*$, Model, $\beta_k$, $b$. Output: $p_{F,\epsilon}(b)$.*

(1) *Partition $\mathcal{D}$ into $M$ disjoint subsets, $(\mathcal{D}_1, \ldots, \mathcal{D}_M)$.*
(2) *In each $\mathcal{D}_j$, where $j = 1, \ldots, M$, fit the Bayesian regression specified in the input Model, and estimate the posterior probability $p_j(b) = \Pr(\beta_k \leq b|\mathcal{D}_j)$.*
(3) *For $j = 1, \ldots, M$, set $\tilde{p}_j(b) = p_j(b)$ when $-2\log(p_j(b)) \leq -2\log(p^*)$, and $\tilde{p}_j(b) = p^*$ otherwise.*
(4) *Draw $\eta$ from a Laplace distribution with scale parameter $-2\log(p^*)/(M\epsilon)$.*
(5) *Compute $F_\epsilon = -2\sum_{j=1}^{M}\log(\tilde{p}_j(b)) + \eta$.*
(6) *Compute $p_{F,\epsilon}(b)$ by referring $F_\epsilon$ to a Monte Carlo estimate of its reference distribution.*

## 3. Evaluation

We use simulations to evaluate the repeated sampling properties of the approaches in Section 2. We begin in Section 3.1 by evaluating the approaches absent privacy concerns; that is, we let $\epsilon = \infty$ so that we add no noise to the aggregated statistics. We then consider the impact of adding noise in Section 3.2.

For both evaluations, we generate 1000 simulated datasets $\mathcal{D}$ of size $n_{\mathcal{D}}$. We examine scenarios with $n_{\mathcal{D}} = 10000$, $n_{\mathcal{D}} = 100000$, and $n_{\mathcal{D}} = 1000000$. For each record $i$, we generate $x_i \sim \mathcal{N}(0,1)$ and $y_i = 1 + 2x_i + \omega_i$, $\omega_i \sim \mathcal{N}(0,1)$. For partitioning, we consider $M \in \{10, 50, 100\}$ samples. We use a Zellner's $g$-prior distribution with $g$ set to the number of rows in the dataset. The posterior distribution of $\beta$ given $\mathcal{D}$ or $\mathcal{D}_j$ (conditional on $\sigma^2$) is thus

$$p(\boldsymbol{\beta}|\sigma^2, \mathcal{D}) \sim \mathcal{N}\left(\frac{n_{\mathcal{D}}}{n_{\mathcal{D}}+1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \frac{n_{\mathcal{D}}}{n_{\mathcal{D}}+1}\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) \tag{3.1}$$

$$p(\boldsymbol{\beta}|\sigma^2, \mathcal{D}_j) \sim \mathcal{N}\left(\frac{n_{\mathcal{D}_j}}{n_{\mathcal{D}_j}+1}(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^T\boldsymbol{y}_j, \frac{n_{\mathcal{D}_j}}{n_{\mathcal{D}_j}+1}\sigma^2(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\right). \tag{3.2}$$

For each simulation run, we compute $p(b)$ using $\mathcal{D}$, which we call the true posterior probability, for all values of $b$ between 1.989 and 2.01 at intervals of .001. We use this sequence because the standard error of the slope $\hat{\beta}_1$ when $n_{\mathcal{D}} = 1000000$ is approximately .001, so that the values of $(b-2)$ here represent number of standard deviations $b$ is from the true $\beta_1 = 2$ in the case with $n_{\mathcal{D}} = 1000000$. We also compute the differentially private posterior probabilities found using $\epsilon = .1$ and $\epsilon = 1$. For the Fisher-derived approach, we use $p^* = 0.001$ and approximate the reference distributions using 10000 Monte Carlo draws.

### 3.1. Results absent privacy considerations.

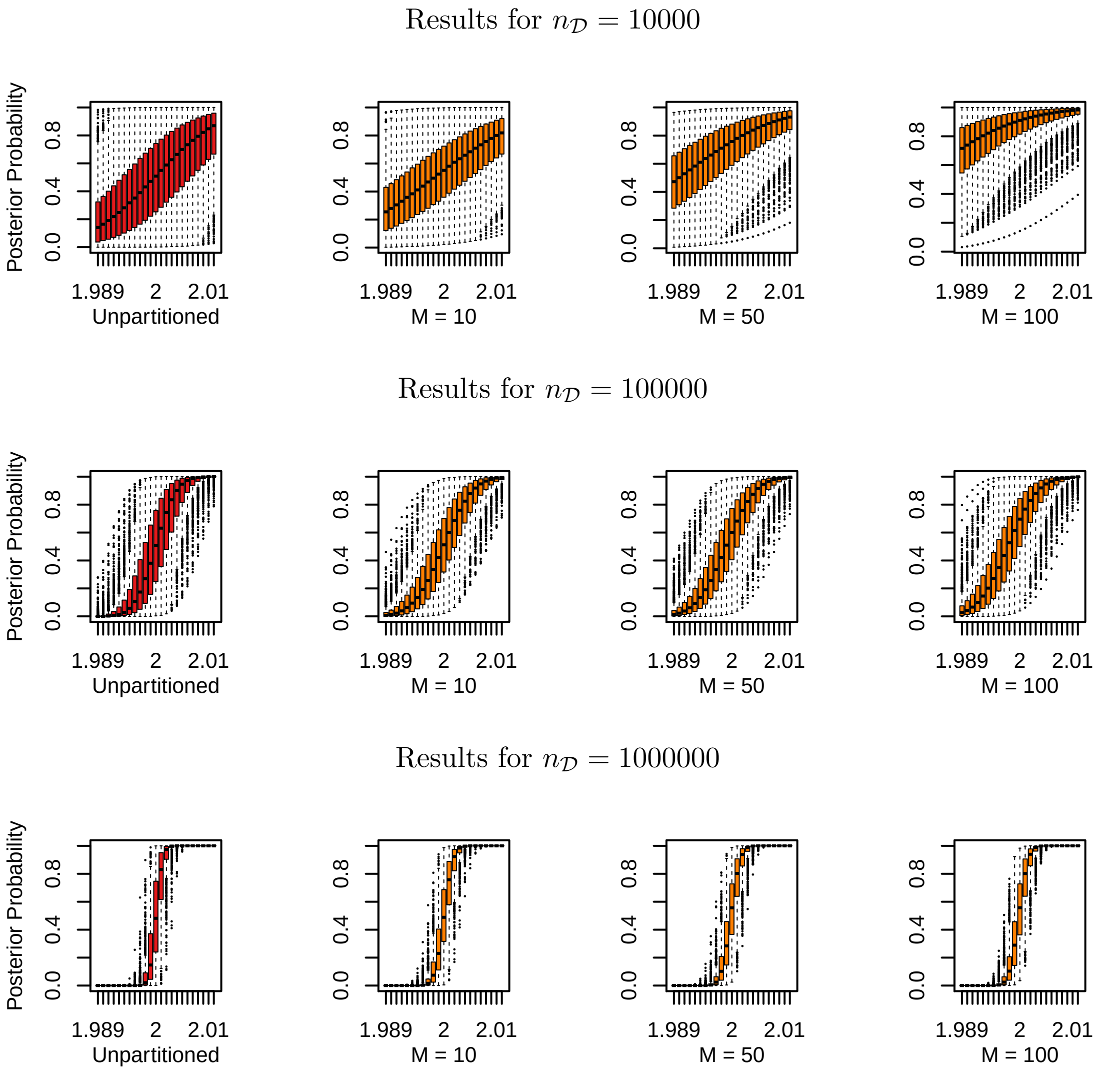
**Results absent privacy considerations.** Absent privacy considerations, ideally, we find that $p_{N,\infty}(b) \approx p(b)$ and $p_{F,\infty}(b) \approx p(b)$ in any simulation run; that is, without privacy the approaches reproduce $p(b)$. Due to the approximations, this is unlikely to be the case generally. More realistically, we would like the repeated sampling distributions of the three quantities to be reasonably similar. That is, if we take repeated samples of size $n$ from the population and apply the approaches to each sample, we want the distributions $f(p_{N,\infty}(b)) \approx f(p(b))$ and $f(p_{F,\infty}(b)) \approx f(p(b))$. Finally, if the sampling distributions are not identical, we would like the expectations to be similar under repeated samples from the population; that is, we want

$$\int p_{N,\infty}(b)d\mathcal{D} \approx \int p(b)d\mathcal{D} \tag{3.3}$$

$$\int p_{F,\infty}(b)d\mathcal{D} \approx \int p(b)d\mathcal{D}. \tag{3.4}$$

These criteria are related to calibrated Bayesian inference (Rubin [1984]).

Figure 1 displays box plots of the distributions of $p(b)$ and of $p_{N,\infty}(b)$ over the 1000 simulation runs. For any box plot in this figure, the line inside the box represents the median value of $p(b)$ or $p_{N,\infty}(b)$ over the 1000 simulation runs; the lines on the lower and upper edges of the box represent the 25th and 75th percentiles over the 1000 simulation runs; the dashed line extending below the 25th percentile ends at the value closest to the first quartile minus 1.5 times the interquartile range; the dashed line extending above the 75th percentile ends at the value closest to the third quartile plus 1.5 times the interquartile range; and, any points extending beyond the dashed lines are values of $p(b)$ or $p_{N,\infty}(b)$ for individual simulation runs. Throughout the article, box plots for other quantities should be interpreted similarly.


Results for $n_{\mathcal{D}} = 10000$

Results for $n_{\mathcal{D}} = 100000$

Results for $n_{\mathcal{D}} = 1000000$

Figure 1: Simulated sampling distributions of the posterior probabilities $p(b)$ and $p_{N,\infty}(b)$ for different values of $M$ and $n_{\mathcal{D}}$ across a range of $b$.

When $n_{\mathcal{D}} \in \{100000, 1000000\}$, we see that the sampling distributions of $p(b)$ and of $p_{N,\infty}(b)$ are similar. This pattern holds across the values of $M$. However, when $n_{\mathcal{D}} = 10000$, the sampling distributions of $p(b)$ and $p_{N,\infty}(b)$ are quite different. In particular, for the runs with $M > 10$, the values of $p_{N,\infty}(b)$ tend to be much larger than the values of $p(b)$, to the point where results are completely unreliable when $M = 100$ for $n_{\mathcal{D}} = 10000$. With large $M$, the sample sizes within each partition become small enough that the approximations used to derive the rescaled-normal approach are not reasonable. Overall, these results suggest that using the rescaled-normal approach is not likely to be reliable unless $n_{\mathcal{D}}$ is large.

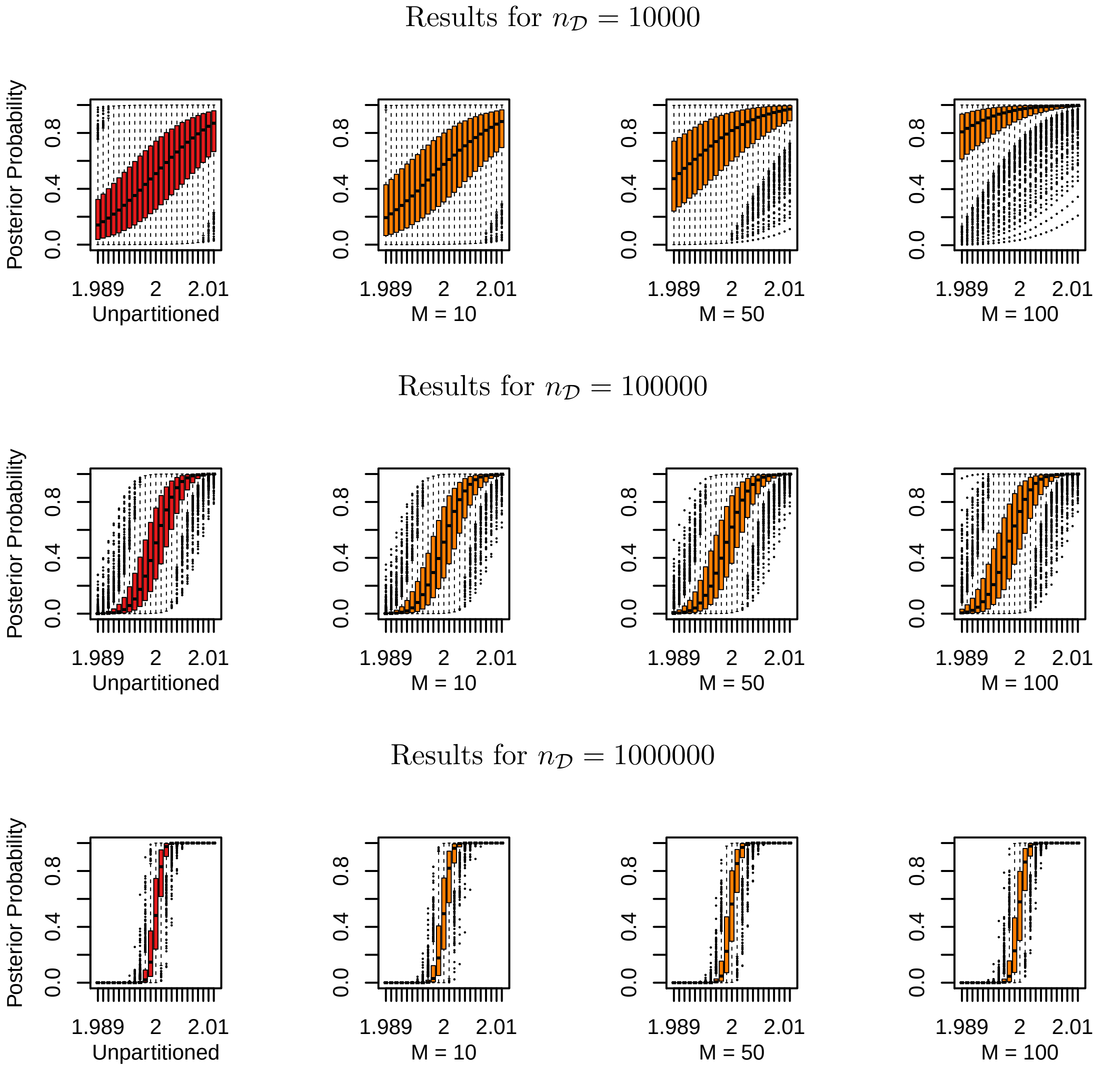


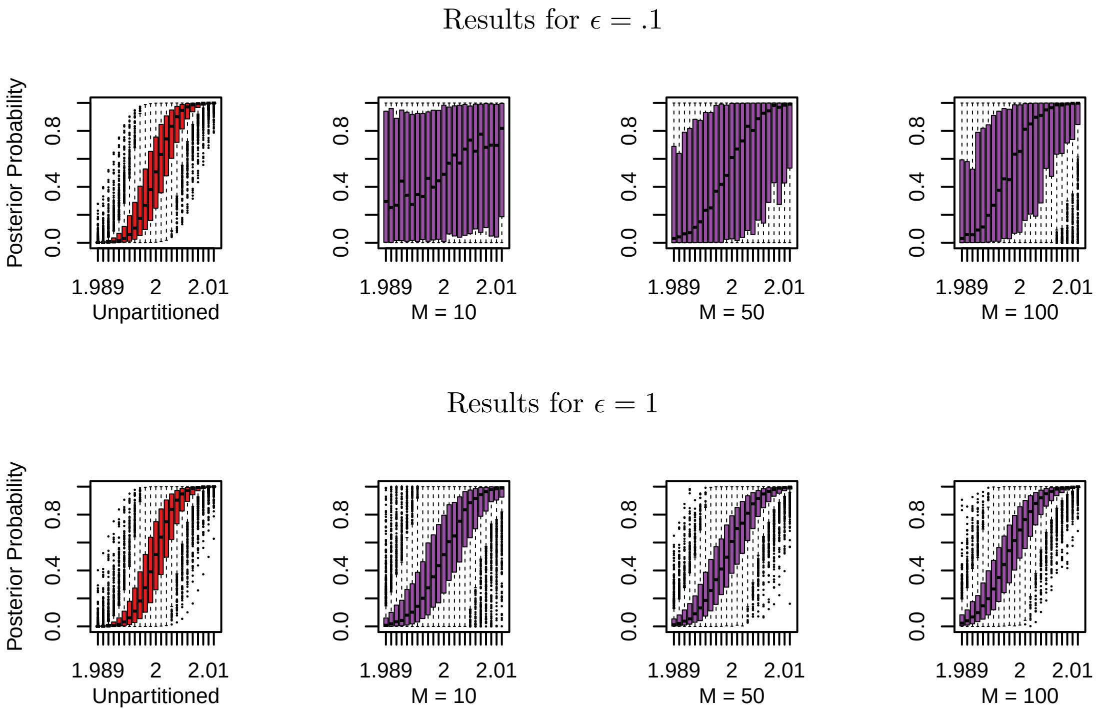
Figure 2: Simulated sampling distributions of the posterior probabilities $p(b)$ and $p_{F,\infty}(b)$ for different values of $M$ and $n_{\mathcal{D}}$ across a range of $b$.

Figure 2 displays the sampling distributions of $p(b)$ and of $p_{F,\infty}(b)$ over the 1000 simulation runs. The results are qualitatively similar to the patterns we see for the rescaled-normal approach. When $n_{\mathcal{D}} \in \{100000, 1000000\}$, the sampling distributions of $p(b)$ and of $p_{F,\infty}(b)$ are similar, even across the values of $M$. However, when $n_{\mathcal{D}} = 10000$ and $M > 10$, the sampling distributions of $p(b)$ and $p_{F,\infty}(b)$ are quite different, making the Fisher-derived approach unreliable in these scenarios. Interestingly, when $n_{\mathcal{D}} = 10000$ and $M = 10$, the sampling distribution of $p_{F,\infty}(b)$ looks more similar to $p(b)$ than the sampling distribution of $p_{N,\infty}(b)$ does. Overall, these results suggest that the Fisher-derived approach may be more effective than the rescaled-normal approach, although once again that both approaches require large sample sizes to be most useful.

3.2. **Results for differentially private versions.** We now investigate the performance of $p_{N,\epsilon}(b)$ and $p_{F,\epsilon}(b)$ when $\epsilon = .1$ and $\epsilon = 1$. Here, we show only results for $n = 100000$ and $n = 1000000$. When $n = 10000$, the sampling distributions of $p_{N,\epsilon}(b)$ for $\epsilon = .1$ are almost uniform on $(0, 1)$ for all values of $b$, making the rescaled-normal approach worthless at $n = 10000$ for $\epsilon = .1$. When $\epsilon = 1$ and $n = 10000$, the sampling distributions mimic those in Figure 1, which we already deem not particularly reliable, with some increased variance. Similarly, when $n = 10000$, we find $p_{F,\epsilon}(b)$ to be unreliable when $\epsilon = .1$. When $\epsilon = 1$, the sampling distributions mimic those those in Figure 2 with some increased variance.

Figure 3 and Figure 4 display the sampling distributions of $p_{N,\epsilon}(b)$ when $n_{\mathcal{D}} = 100000$ and $n_{\mathcal{D}} = 1000000$, respectively. For both sample sizes, the rescaled-normal approach provides unreliable results when $\epsilon = .1$, not even matching $p(b)$ in expectation. We note that for these sample sizes, the performance of $p_{N,\epsilon}(b)$ does improve as $M$ gets larger. This mainly reflects the reduced impact of the noise from the Laplace mechanism. When $\epsilon = 1$, the sampling distributions of $p_{N,\epsilon}(b)$ more closely resemble those of $p(b)$, although there remain systematic differences in the distributions particularly when $n = 100000$.

Turning to the Fisher-derived approach, Figure 5 and Figure 6 display results when $n_{\mathcal{D}} = 100000$ and $n_{\mathcal{D}} = 1000000$, respectively. When $\epsilon = 0.1$, the sampling distributions of $p_{F,.1}(b)$ resemble those of $p(b)$ when $M = 50$ and $M = 100$, but not when $M = 10$. Apparently, the subsample and aggregate method injects enough noise to degrade the performance of $p_{F,.1}(b)$ at $M = 10$. In contrast, when $\epsilon = 1$, $p_{F,1}(b)$ performs reasonably well even when $M = 10$. We note that sampling distribution of $p_{F,1}(b)$ resembles that of $p(b)$ at all values of $M$.



Results for $\epsilon = .1$

Results for $\epsilon = 1$

Figure 3: Simulated sampling distributions of the posterior probabilities $p(b)$ and $p_{N,\epsilon}(b)$ when $n_{\mathcal{D}} = 100000$ for $\epsilon \in \{.1, 1\}$ and different values of $M$ across a range of $b$.

## 4. Differentially Private Posterior Quantiles

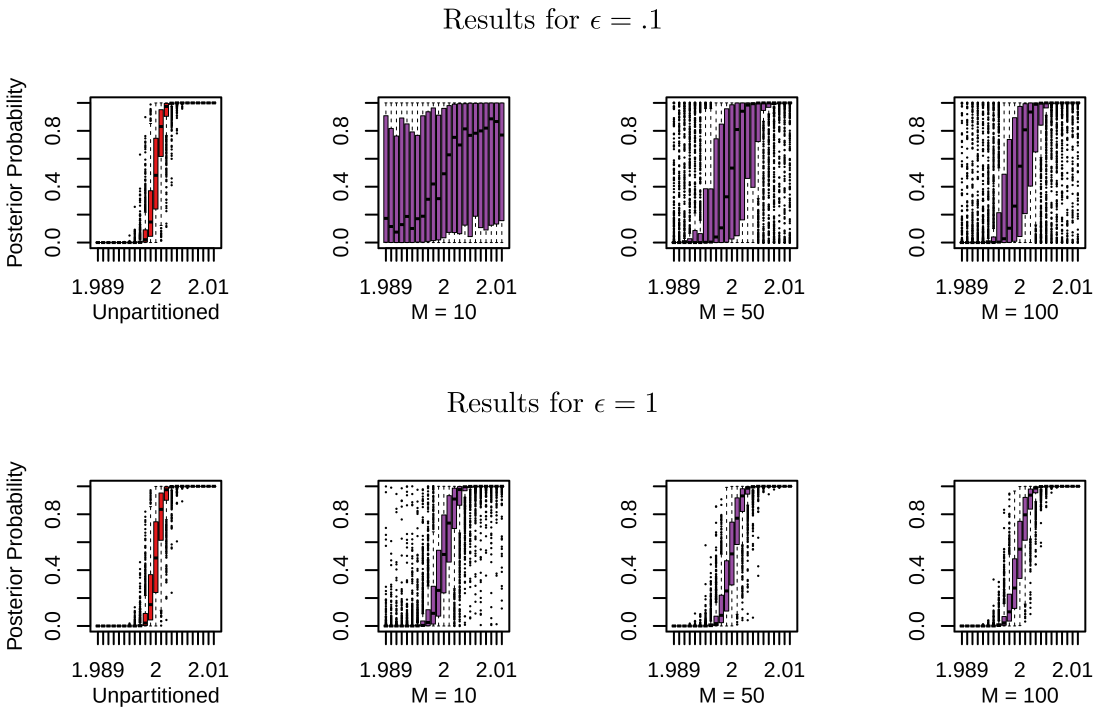
In this section, we outline a differentially private approach for obtaining posterior quantiles of parameters in Bayesian inference. For regression contexts, we seek the posterior quantile of some $\beta_k$, which we write as $Q(p) = \{b : \Pr(\beta_k \leq b|\mathcal{D}) = p\}$.

Our approach relies on the methods developed by Li et al. [2017], who present an approach for finding posterior quantiles in large datasets. Loosely speaking, they partition $\mathcal{D}$ into $M$ disjoint subsets, compute the posterior quantile of interest using each $\mathcal{D}_j$, and average the $M$ estimated posterior quantiles to derive an estimate of $Q(p)$. More precisely, in each $\mathcal{D}_j$, they find a rescaled posterior distribution,

$$p^*(\boldsymbol{\beta}, \sigma^2|\mathcal{D}_j) \propto \left[\prod_{i \in \mathcal{D}_j} p(y_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_i)\right]^M p(\boldsymbol{\beta}, \sigma^2), \tag{4.1}$$

where $p(\boldsymbol{\beta}, \sigma^2)$ is the prior distribution that the analyst would use for an analysis of $\mathcal{D}$. The factor of $M$ in the likelihood serves to rescale the variance of the posterior distribution based on any partition to match that of the posterior distribution based on the full dataset. Li et al. [2017] prove that this approach results in accurate estimates of $Q(p)$, absent privacy concerns.

This approach naturally suggests an application of subsample and aggregate. Let $Q_j^*(p)$ be the value of $b$ such that $\Pr^*(\beta \leq b|\mathcal{D}_j) = p$, where the density function used to compute


Results for $\epsilon = .1$

Results for $\epsilon = 1$

Figure 4: Simulated sampling distributions of the posterior probabilities $p(b)$ and $p_{N,\epsilon}(b)$ when $n_{\mathcal{D}} = 1000000$ for $\epsilon \in \{.1, 1\}$ and different values of $M$ across a range of $b$.

the quantile is given in (4.1). We would like to set $h(\mathcal{D}_j) = Q^*_j(p)$ and

$$g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) = \sum_{j=1}^{M} Q^*_j(p)/M. \tag{4.2}$$

To ensure separated posterior distributions in each $\mathcal{D}_j$, which is useful for differential privacy, we require the prior distribution for $(\boldsymbol{\beta}, \sigma^2)$ not be data-dependent.

We cannot do so directly, however. Posterior quantiles are unbounded and thus have indeterminate global sensitivity. We therefore need to put bounds on the values of the terms used in (4.2). We propose that analysts allocate some of the privacy budget, say $\epsilon_1 < \epsilon$, to find noisy bounds on the $M$ values of $h(\mathcal{D}_j)$, and use the remainder of the budget to add noise to $g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M))$ via a Laplace mechanism.

An example of a bounding algorithm is given by Chen et al. [2018]. Their algorithm takes a list of numbers as an input. It returns a noisy interval centered around zero that contains approximately $\theta$ percent of all the values in the list, where the analyst specifies $\theta$ (e.g., $\theta = 95$). With such bounds, say $(-c, c)$, analysts can truncate each $Q^*_j(p)$ at the bounds before entering them in (4.2). The global sensitivity of the mean of the clipped quantiles is then $2c/M$. The remaining privacy budget $\epsilon_2 = \epsilon - \epsilon_1$ is used in the Laplace mechanism to add noise to the average.

We did some preliminary investigations with the bounding algorithm of Chen et al. [2018] and found it can result in wide bounds for the regression quantiles. The algorithm provides

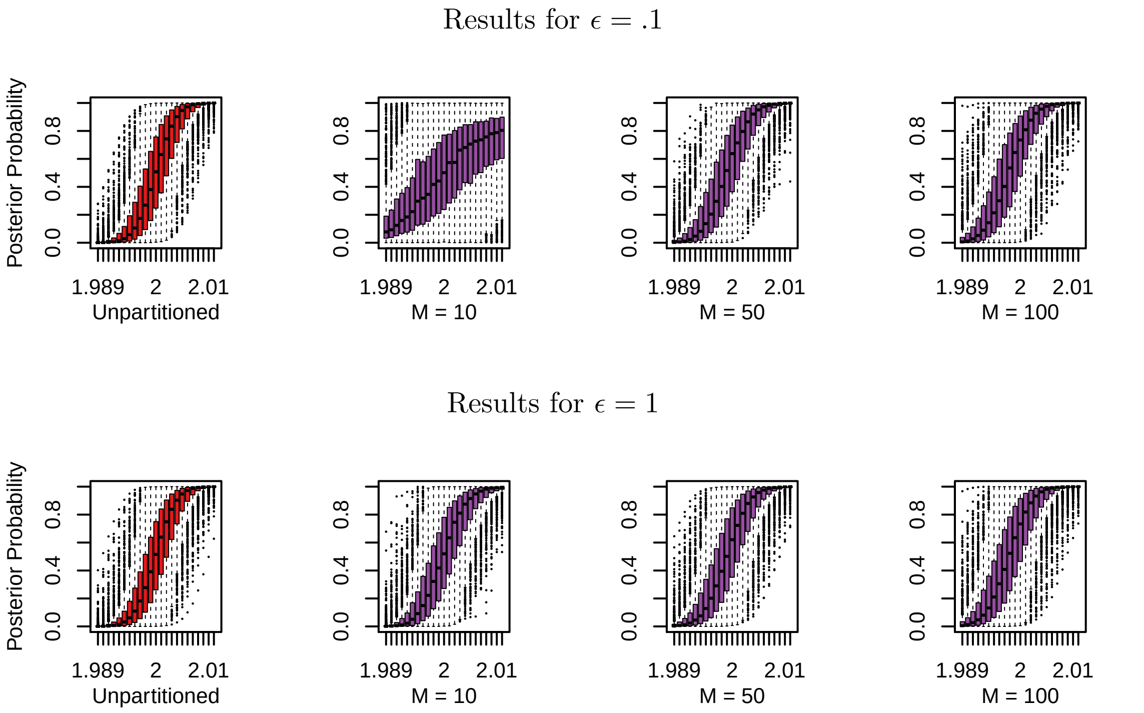
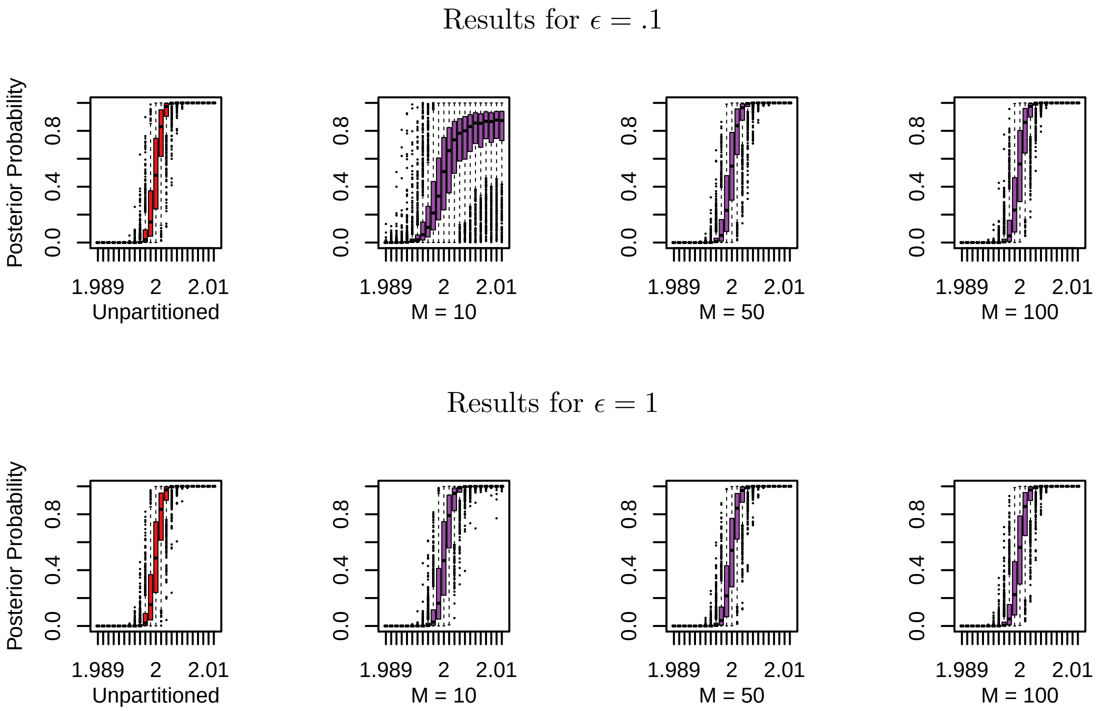


Figure 5: Simulated sampling distributions of the posterior probabilities $p(b)$ and $p_{F,\epsilon}(b)$ when $n_{\mathcal{D}} = 100000$ for $\epsilon \in \{.1, 1\}$ and different values of $M$ across a range of $b$.

bounds that are centered around zero, whereas the values of $Q_j^*(p)$ are not necessarily centered around zero. As a result, the application of subsample and aggregate can inject too much noise for reasonable privacy budgets. To get around this problem, one approach is to further allocate the privacy budget to find the center of the $M$ values of $Q_j^*(p)$; for example, use smooth sensitivity to estimate the median of the $M$ values. Analysts can apply the method of Chen et al. [2018] to the median-centered values of $Q_j^*(p)$. This should produce tighter bounds, which should result in less noise in the Laplace mechanism used in subsample and aggregate. We plan to investigate the performance of this algorithm in future work.

## 5. Conclusion

The empirical investigations here are limited in scope, and additional research is needed. However, the results of the simulations suggest some general conclusions. First, the rescaled-normal approach does not work well with $\epsilon = .1$ for sample sizes up to 1000000, nor does it work well when sample sizes are on the order of 10000. In short, the simulations do not support extensive use of the rescaled-normal approach for differentially private Bayesian inference. On the other hand, the Fisher-derived approach seems to perform well, offering reasonable performance even for $\epsilon = .1$ when sample sizes exceed 100000. At the moment, the justification for the Fisher-derived approach is *ad hoc*. In future work, we intend to explore theoretically when one can expect the Fisher-derived approach to be effective.


Results for $\epsilon = .1$

Results for $\epsilon = 1$

Figure 6: Simulated sampling distributions of the posterior probabilities $p(b)$ and $p_{F,\epsilon}(b)$ when $n_{\mathcal{D}} = 1000000$ for $\epsilon \in \{.1, 1\}$ and different values of $M$ across a range of $b$.

In any particular application, the performances of the algorithms depend on the appropriateness of the approximations. These approximations are difficult to evaluate in ways that still satisfy differential privacy, as the most direct checks require access to confidential data values. To enable evaluation, it may be possible to adapt the strategy of Barrientos et al. [2018]. They use subsample and aggreate techniques to develop differentially private measures for verifying that estimated regression coefficients fall within pre-specified tolerance levels. Developing such measures is a subject of future work.

In some contexts, an alternative approach for differentially private Bayesian inference involves (i) adding differentially private noise to sufficient statistics for the model of interest and (ii) determining the posterior distribution of model parameters given the noisy statistics [e.g., Charest, 2010, Williams and Mcsherry, 2010, Karwa et al., 2015]. When feasible, this approach has an advantage over our subsample and aggregate methods, in that one need not rely on the validity of the large-sample approximations we make. However, such approaches can be computationally very intensive, particularly for models that have a large number of sufficient statistics. Additionally, the sufficient statistics for the model could have high sensitivity, resulting in substantial loss of accuracy from basing posterior distributions on noisy quantities. Nonetheless, it would be interesting to compare the performance of such approaches to the performance of the algorithms presented here.

## References

A. F. Barrientos, A. Bolton, T. Balmat, J. P. Reiter, J. M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, and M. DeLong. Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *The Annals of Applied Statistics*, 12:1124–1156, 2018.

A. S. Charest. How can we analyze differentially private synthetic datasets. *Journal of Privacy and Confidentiality*, 2:2:Article 3, 2010.

Y. Chen, A. F. Barrientos, A. Machanavajjhala, and J. P. Reiter. Is my model any good: differentially private regression diagnostics. *Knowledge and Information Systems*, 54(1): 33–64, Jan 2018. doi: 10.1007/s10115-017-1128-z.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, Aug. 2014. doi: 10.1561/0400000042.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

R. Fisher. *Statistical Methods for Research Workers.* Edinburgh Oliver & Boyd, 1925.

P. D. Hoff. *A First Course in Bayesian Statistical Methods.* Springer Publishing Company, Inc., 2009.

V. Karwa, D. Kifer, and A. B. Slavkovic. Private posterior distributions from variational approximations. *CoRR*, abs/1511.07896, 2015. URL http://arxiv.org/abs/1511.07896.

C. Li, S. Srivastava, and D. B. Dunson. Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104(3):665–680, 2017. doi: 10.1093/biomet/asx033.

K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, STOC '07, pages 75–84, New York, NY, USA, 2007. ACM. doi: 10.1145/1250790.1250803.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984. doi: 10.1214/aos/1176346785.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley, 1987.

O. Sheffet. Differentially private least squares: Estimation, confidence and rejecting the null hypothesis. *CoRR*, 2015. URL http://arxiv.org/abs/1507.02482.

O. Williams and F. Mcsherry. Probabilistic inference and differential privacy. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2451–2459. Curran Associates, Inc., 2010. URL http://papers.nips.cc/paper/3897-probabilistic-inference-and-differential-privacy.pdf.