NSERC Summer Project

Rob Li

Supervised by Prof. Mehdi Dagdoug
and Hui Shen

August 17, 2025

Abstract

This project investigates differentially private estimation in the survey sampling setting, focusing on the Generalized Regression (GREG) estimator for Simple Random Sampling Without Replacement (SRSWOR). We formalize sample-level and population-level notions of privacy, and construct three candidate private GREG estimators with confidence intervals: NoisyStats, NoisyModel, and NoisySample. Through theoretical analysis and Monte Carlo simulations, we find that while NoisyModel and NoisySample suffer from excessive variance, NoisyStats remains a viable approach. The methods and findings presented here are a proof-of-concept, intended as groundwork for subsequent extensions to higher dimensions, alternative sampling designs, and real data applications

Contents

1 Introduction

Survey organizations face a dual mandate: publish accurate finite-population statistics while safeguarding the privacy of contributing individuals. Differential Privacy (DP) is a principled standard for this purpose, and survey sampling is a natural but underdeveloped arena for its application—despite its central role in public policy and the frequent public release of survey statistics. Existing DP work in this area has largely targeted design-based estimators like Horvitz–Thompson, leaving a gap for *model-assisted* methods that are pervasive in practice. In particular, there is no established differentially private version of the linear-regression–based GREG estimator, even though it is among the most commonly used tools for leveraging auxiliary frames to improve precision.

This project seeks to begin addressing that gap. We study DP estimation of a finite-population mean in the model-assisted framework under standard sampling designs, with a focus on SRSWOR and public auxiliary frames. Guided by a brief literature review, we adapt neighboring definitions to the survey setting, define differentially private model-assisted estimators, and evaluate candidate constructions both theoretically and empirically before studying their properties.

2 Background

This section reviews background, terminology, and notation for Differential Privacy. We provide a formulation for its application in Survey Sampling, and introduce model assisted estimators including the Generalized Regression Estimator (GREG).

2.1 Differential Privacy

Differential Privacy (DP) is a mathematical framework that quantifies the privacy guarantee when releasing sensitive information like statistical queries from datasets. At its core, DP ensures that the presence or absence of any individual in the dataset has a limit impact on the releases output.

In the DP setting, we assume that a **database universe**, $\mathscr{D}$ is a set consisting of all the possible datasets $\mathcal{D}$ we could have. Elements of the dataset $d_i \in \mathcal{D}$ are individuals, and $\mathcal{D}_0$, the super dataset, is the set of all individuals that could be in a dataset. It is possible that $\mathcal{D}_0 \notin \mathscr{D}$ such as in contexts where all possible datasets are of a fixed size; this is know as *bounded differential privacy.* Contexts where possible datasets range in size are under *unbounded differential privacy.* The discussion of the effect of *the presence or absence any individual in the dataset* is formalized with the concept of **neighboring datasets**

Definition 2.1 (neighboring datasets under bounded DP) *Datasets $\mathcal{D} \in \mathscr{D}$, and $\mathcal{D}' \in \mathscr{D}$ are* neighbors *if $\mathcal{D} \backslash \mathcal{D}' = \{i\}$, and $\mathcal{D}' \backslash \mathcal{D} = \{j\}$ for some $i \neq j \in \mathcal{D}_0$.*

In other words, neighboring datasets in this setting are characterized by replacing one individual or another. For unbounded DP, neighboring datasets are characterized by the addition or removal of an individual.

Definition 2.2 (neighboring datasets under unbounded DP) *Datasets $\mathcal{D} \in \mathscr{D}$, and $\mathcal{D}' \in \mathscr{D}$ with $|\mathcal{D}| = |\mathcal{D}'| + 1$ are* neighbors *if $\mathcal{D} \backslash \mathcal{D}' = \{i\}$ for some $i \in \mathcal{D}_0$.*

It is worth observing that any pair of neighbors in bounded DP are two-doors-down from each other in unbounded DP.

Beyond characterizing differing datasets, DP is focused on measuring and limiting the impact of slight changes in the dataset on specific *statistical query* functions $q : \mathscr{D} \to \mathcal{Y}$. *Differentially private mechanisms* $\tilde{q}(\cdot)$ are always (with the exception of constant response) randomized algorithms, so we measure the difference between $\Pr[\widetilde{q}(\mathcal{D}) = t]$ and $\Pr[\widetilde{q}(\mathcal{D}') = t]$.

A standard configuration of DP is **approximate differential privacy**, or (ε, δ)-DP. It measures the influence of

Definition 2.3 ((ε, δ)-Differential Privacy) *A randomized mechanism $\widetilde{q} : \mathscr{U} \to \mathcal{Y}$ satisfies* (ε, δ)-differential privacy *if for all measurable sets $S \subseteq \mathcal{Y}$ and for all neighboring datasets $D, D' \in \mathscr{U}$ differing in at most one element,*

$$\Pr[\widetilde{q}(D) \in S] \leq e^{\varepsilon} \Pr[\widetilde{q}(D') \in S] + \delta.$$

When $\delta = 0$, the guarantee is called ε-differential privacy*.*

An alternative formulation of DP is *zero-concentrated differential privacy (zCDP)* which offers convent composition proprieties and a tighter connection to Additive Gaussian Noise mechanisms.

Definition 2.4 (ρ-Zero-Concentrated Differential Privacy) *A randomized mechanism $\widetilde{q}$ satisfies* ρ-zCDP *if for all neighboring datasets $D, D' \in \mathscr{U}$ and all $\alpha \in (1, \infty)$,*

$$D_\alpha(\widetilde{q}(D) \,\|\, \widetilde{q}(D')) \leq \rho\alpha,$$

where $D_\alpha(\cdot\|\cdot)$ denotes the order-α Rényi divergence.

We will do out analysis under zCDP, but the following gives a standard conversion from zCDP to approximate DP.

Proposition 2.1 (Conversion from zCDP to (ε, δ)-DP) *If a mechanism $\widetilde{q}$ satisfies ρ-zCDP, then for all $\delta \in (0, 1)$ it also satisfies (ε, δ)-differential privacy with*

$$\varepsilon = \rho + 2\sqrt{\rho \log(1/\delta)}.$$

Equivalently, for any $\varepsilon \geq \rho$,

$$\delta \leq \exp\left(-\frac{(\varepsilon - \rho)^2}{4\rho}\right).$$

The DP mechanism we will be studying is configured based on a sensitivity analysis. This analysis looks at the maximum difference between $\widetilde{q}(\mathcal{D})$ and $\widetilde{q}(\mathcal{D}')$.

Definition 2.5 (Sensitivity) *For a function $q : \mathscr{U} \to \mathbb{R}$, the (global) ℓ_2-sensitivity is*

$$\Delta(q) := \max_{D, D' \text{ neighbors}} \|q(D) - q(D')\|_2 .$$

With a sensitivity analysis, we can configure often used *Gaussian Mechanism* for differential privacy.

Proposition 2.2 (Gaussian Mechanism for zCDP) *Let $q : \mathscr{U} \to \mathbb{R}$ have ℓ_2-sensitivity $\Delta(q)$. Consider the mechanism $\widetilde{q}$ that on input D releases*

$$\widetilde{q}(D) = q(D) + \mathcal{N}\left(0, \frac{\Delta(q)^2}{2\rho}\right).$$

Then $\widetilde{q}$ satisfies ρ-zCDP.

Proposition 2.3 (Composition) *Let $\widetilde{q}_1$ satisfy ρ_1-zCDP and $\widetilde{q}_2$ satisfy ρ_2-zCDP. Then the composed mechanism $(\widetilde{q}_1, \widetilde{q}_2)$ satisfies $(\rho_1 + \rho_2)$-zCDP.*

The following are some important properties of differentially private mechanism.

Proposition 2.4 (Post-processing) *If $\widetilde{q}$ satisfies ρ-zCDP and f is any (possibly randomized) mapping independent of the dataset, then $f \circ \widetilde{q}$ also satisfies ρ-zCDP.*

Theorem 2.5 (Privacy Amplification by Subsampling) *Let $\widetilde{q} : \mathscr{D} \to \mathcal{Y}$ be an (ε, δ)-differentially private mechanism. Let $\mathcal{S} \subset \mathcal{D}$ be a random sample selected by a sampling design $P(\cdot|\mathcal{D}))$. Define $\widetilde{\widetilde{q}}(\mathcal{D}) := \widetilde{q}(\mathcal{S})$. If*

1. *If $\mathcal{D}, \mathcal{D}' \in \mathscr{D}$ are neighbors then their exists respective subsamples $\mathcal{S} \subset \mathcal{D}$ and $\mathcal{S}' \subset \mathcal{D}$ which are also neighboring datasets (differing at the individuals $\mathcal{D}$ and $\mathcal{D}'$ differ).*
2. *If $\mathcal{D}, \mathcal{D}' \in \mathscr{D}$ are neighbors, and d_k is in both datasets, then*
   - $\Pr[d_k \in \mathcal{S} | i \in \mathcal{S}] = \Pr[d_k \in \mathcal{S}' | j \in \mathcal{S}']$ *in the bounded setting*
   - $\Pr[d_k \in \mathcal{S} | i \in \mathcal{S}] = \Pr[d_k \in \mathcal{S}' | \mathcal{S}']$ *in the unbounded setting*

Then $\widetilde{\widetilde{q}}$ is (ε', δ')-differentially private with:

$$\varepsilon' = \log(1 + \pi(e^\varepsilon - 1)), \quad \delta' = \pi_i \delta$$

where $\pi := \max_{k \in \mathcal{D}_0} \Pr[d_k \in S']$ is the sampling probability.

2.2 Differentially Private Sampling

In this project, we were interested in differential privacy in the sampling theory context to eventually construct robust differentially private model assisted estimators. In sampling theory, we consider a finite **population** $\mathcal{U} = \{1, 2, ..., N\}$ of size N where each person $k \in \mathcal{U}$ has a value y_k as the **variable of interest**, and possibly one or more **auxiliary variables** x_k. The goal of survey sampling is to estimate a population parameter such as the mean

$$\mu_y = \frac{1}{n} \sum_{k \in \mathcal{U}} y_k$$

using information of the variable of interest collected from a sample $\mathcal{S} \subset \mathcal{U}$, drawn by some **sampling design** $P(\cdot|\mathcal{U})$. One such sampling design is *simple random sampling without replacement (SRSWOR(n))* where the only possible samples are size n subsets of $\mathcal{U}$, and each has an equal probability of selection.

Estimation of the parameter of interest may involve a *design-based* approach, *model-based* approach, or *model assisted estimators*. We will be working privatizing the GREG estimator, a model assisted estimator.

Applying differential privacy to this context introduces another layer of complexity. There are two datasets in the survey sampling setting, the population $\mathcal{U}$, and the sample $\mathcal{S}$. This creates two different levels for privacy which offer different levels of protection.

1. **Sample Level DP** In sample level DP, the dataset we are protecting is $\mathcal{S}$. The *sample universe* is all possible sample based on the design $\mathscr{S} = P(\cdot|\mathcal{U})$. The *super-sample* is the population $\mathcal{U}$ which is fixed or invariant. The values $\{x_k\}_{k \in \mathcal{U}}$, N are public while $\{(x_k, y_k)\}_{k \in \mathcal{S}}$ are sensitive. For fixed sized sampling designs, n may be public.
2. **Population Level DP** In population level DP, the dataset we are protecting is $\mathcal{U}$. The *population universe* $\mathscr{U}$ is a set of all possible populations. The *super-population* $\mathcal{W}$ is the set of individuals that the plausibly be in the population. The values $\{x_k\}_{k \in \mathcal{U}}$, and $\{(x_k, y_k)\}_{k \in \mathcal{S}}$ are all sensitive. If the setting is bounded population DP, then the population size N may be public. As a consequence, this also ensures sample level DP.

2.3 Model Assisted Estimators

In survey sampling we often aim at measuring population parameters such as the mean using only the sample collected. When auxiliary information about each individual $\mathbf{x}_k$ is available, it can be leveraged to construct more efficient **model assisted** estimators.

Model assisted estimators use the sampled data to train a model to guide estimator for the entire population. Inference is still design based since we also consider the randomization from the sampling design rather than just the validity of the model. This makes model assisted estimators more robust that purely model-based approaches while also still achieving variance reduction from the design based methods.

One of the widely used model-assisted estimators is the **Generalized Regression (GREG)** estimator which incorporates the auxiliary variables with ordinary linear regression.

$$\widehat{\mu}_{\text{greg}} = \frac{1}{N} \sum_{k \in \mathcal{U}} \mathbf{x}_k^\top \widehat{\beta} + \frac{1}{N} \sum_{k \in \mathcal{S}} \frac{y_k - \mathbf{x}_k^\top \widehat{\beta}}{\pi_k}, \quad \text{where } \widehat{\beta} = \left(\sum_{k \in \mathcal{S}} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in \mathcal{S}} \mathbf{x}_k y_k,$$

and π_k is the sampling probability for k under the sampling design used. For SRSWOR(n), $\pi_k = n/N$ for all $k \in \mathcal{U}$

For 1-D auxiliary data, we can separate the intercept in the following form.

$$\widehat{\mu}_{\text{greg}} = \frac{1}{N} \sum_{k \in \mathcal{U}} (\widehat{\alpha} + x_k \widehat{\beta}) + \frac{1}{N} \sum_{k \in \mathcal{S}} \frac{y_k - (\widehat{\alpha} + x_k \widehat{\beta})}{\pi_k},$$

$$\text{where } \widehat{\beta} = \left(\sum_{k \in \mathcal{S}} x_k x_k^\top \right)^{-1} \sum_{k \in \mathcal{S}} x_k y_k, \quad \text{and } \widehat{\alpha} = \frac{1}{n} \sum_{k \in \mathcal{S}} y_k - \frac{1}{n} \sum_{k \in \mathcal{S}} x_k \widehat{\beta}$$

Looking back to the augmented model which includes the intercept $\mathbf{x}_k = [1 \quad x_k]^\top$, $\widehat{\beta} = [\widehat{\alpha} \quad \widehat{\beta}_1]^\top$, we can leverage the fact that the sample sum of residuals are zero. This is because, the OLS solves for $\arg\min_{\widehat{\beta}} \sum_{k \in \mathcal{S}} (y_k - \mathbf{x}_k^\top \widehat{\beta})^2$,

$$E := \sum_{k \in \mathcal{S}} (y_k - \mathbf{x}_k^\top \widehat{\beta})^2 = \sum_{k \in \mathcal{S}} (y_k - 1\widehat{\alpha} - x_k^\top \widehat{\beta}_1)^2$$

$$0 =: \frac{dE}{d\widehat{\alpha}} = -2\sum_{k\in\mathcal{S}}(y_k - 1\widehat{\alpha} - x_k^\top\widehat{\beta}_1) = -2\sum_{k\in\mathcal{S}}(y_k - \mathbf{x}_k^\top\widehat{\beta}) \implies \sum_{k\in\mathcal{S}}(y_k - \mathbf{x}_k^\top\widehat{\beta}) = 0$$

Therefore, we can also write

$$\widehat{\mu}_{\text{greg}} = \frac{1}{N}\sum_{k\in\mathcal{U}}\mathbf{x}_k^\top\widehat{\beta} + \frac{1}{N}\sum_{k\in\mathcal{S}}\frac{y_k - \mathbf{x}_k^\top\widehat{\beta}}{\pi_k} = \mu_x^\top\widehat{\beta}.$$

3 Theoretical Sample Level DP GREG

In this section, we formalize three *sample-level* DP constructions for the GREG mean confidence intervals under SRSWOR(n). In this environment, the auxiliary frame $\{x_k\}_{k\in U}$ (hence μ_x) and N are public, while $\{(x_k, y_k)\}_{k\in S}$ are sensitive. Throughout, we work under ρ-zCDP, and used Gaussian mechanisms calibrated from global ℓ_2 sensitivities after an a priori bounds $x_k \in [-B_x, B_x]$ and $y_k \in [-B_y, B_y]$, and we compose privacy budgets across releases.

Algorithm 1 (NoisyStats) privatizes the sufficient statistics $\bar{x}, \bar{y}, \overline{xx}, \overline{xy}, \overline{yy}$ and calculates a centred GREG ; its variance estimator is likewise privatized, and the extra variance from injected noise is approximated via a delta-method gradient in those stats.

Algorithm 2 (NoisyModel) seeks to inject noise directly into $\hat{\beta}$ (and $\hat{V}_{\text{greg}}$) using a leave-one-out–based sensitivity bound, yielding an explicit noise contribution $\mu_x^2\sigma_\beta^2$ in the variance.

Algorithm 3 (NoisySample) releases a synthetic privatized sample $(\tilde{x}_k, \tilde{y}_k)_{k\in S}$ via a vector mechanism and then computes the usual GREG point and variance estimates on that synthetic data, with DP guaranteed by post-processing.

These three designs trade off analytical simplicity, tightness of sensitivity (hence noise scale), and ease of extension to higher dimensions; we benchmark them side-by-side in the simulations that follow.

3.1 Algorithm 1: NoisyStats 1D

Private Point Estimate Construction In this setting, there is only one axillary variable x_k. We will be working with the centred form for $\widehat{\beta}$. The GREG estimator for SRSWOR can be rewritten by

$$\begin{aligned}
\widehat{\mu}_{\text{greg}} &= \frac{1}{N}\sum_{k\in\mathcal{U}}(\widehat{\alpha} + x_k\widehat{\beta}) + \frac{1}{N}\sum_{k\in\mathcal{S}}\frac{y_k - \widehat{\alpha} - x_k\widehat{\beta}}{\pi_k} \\
&= \widehat{\alpha} + \frac{1}{N}\sum_{k\in\mathcal{U}}x_k\widehat{\beta} + \sum_{k\in\mathcal{S}}\frac{y_k - \widehat{\alpha} - x_k\widehat{\beta}}{n} \\
&= \widehat{\alpha} + \frac{1}{N}\sum_{k\in\mathcal{U}}x_k\widehat{\beta} + \frac{1}{n}\sum_{k\in\mathcal{S}}y_k - \widehat{\alpha} - \frac{1}{n}\sum_{k\in\mathcal{S}}x_k\widehat{\beta} \\
&= \mu_x\widehat{\beta} + \overline{y} - \overline{x}\widehat{\beta} = (\overline{y} - \overline{x}\widehat{\beta}) + \mu_x\widehat{\beta} = \widehat{\alpha} + \mu_x\widehat{\beta}
\end{aligned}$$

$$\widehat{\alpha} = \overline{y} - \overline{x}\widehat{\beta}, \quad \widehat{\beta} = \frac{\sum_{k\in\mathcal{S}}x_k y_k - \sum_{k\in\mathcal{S}}x_k\sum_{k\in\mathcal{S}}y_k}{\sum_{k\in\mathcal{S}}x_k x_k - \sum_{k\in\mathcal{S}}x_k\sum_{k\in\mathcal{S}}x_k} = \frac{\overline{xy} - \overline{x}(\overline{y})}{\overline{xx} - \overline{x}(\overline{x})}$$

The sensitive parts of $\widehat{\mu}_{\text{greg}}$ in this form are reduced to the sufficient statistics

$$\overline{x} = \frac{1}{n}\sum_{k\in\mathcal{S}}x_k, \quad \overline{y} = \frac{1}{n}\sum_{k\in\mathcal{S}}y_k, \quad \overline{xy} = \frac{1}{n}\sum_{k\in\mathcal{S}}x_k y_k, \quad \overline{xx} = \frac{1}{n}\sum_{k\in\mathcal{S}}x_k^2.$$

The idea of this first algorithm is the privatize these sufficient statistics with Gaussian noise and then compose the privatized versions to construct a private $\widehat{\mu}_{\text{greg}}$ estimator.

Private Variance Estimate Construction In addition to construct confidence intervals, we will have to privatize the variance estimation for SRSWOR $\widehat{\mu}_{\text{greg}}$. We can write $\hat{V}_{\text{greg}}$ in terms of the sufficient statistics above along with $\overline{y} - \frac{1}{n}\sum_{k\in\mathcal{S}}y_k^2$. The standard form for the unbiased variance for SRSWOR is

$$\hat{V} = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_{k\in\mathcal{S}}(y_k - \overline{y})^2$$

For a GREG estimator, we replace the y_k's in the equation with the residuals from the model $\hat{\epsilon}_k = y_k - \widehat{\alpha} - x_k\widehat{\beta}$

$$\begin{aligned}\hat{V}_{\text{greg}} &= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{k\in\mathcal{S}}(\hat{\epsilon}_k - \bar{\hat{\epsilon}})^2 \\ &= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left(\sum_{k\in\mathcal{S}}\hat{\epsilon}_k^2 - n\bar{\hat{\epsilon}}^2\right).\end{aligned}$$

We have

$$\begin{aligned}\sum_{k\in\mathcal{S}}\hat{\epsilon}^2 &= \sum_{k\in\mathcal{S}}(y_k - \widehat{\alpha} - x_k\widehat{\beta})^2 \\ &= \sum_{k\in\mathcal{S}} y_k^2 - 2\widehat{\beta}\sum_{k\in\mathcal{S}} x_k y_k + \widehat{\beta}\left(\sum_{k\in\mathcal{S}} x_k x_k\right)\widehat{\beta} + \widehat{\alpha}^2 - 2\widehat{\alpha}\sum_{k\in\mathcal{S}} y_k + 2\widehat{\alpha}\sum_{k\in\mathcal{S}} x_k\widehat{\beta} \\ &= n(\overline{yy} - 2\widehat{\beta}\overline{xy} + \widehat{\beta}\overline{xx}\widehat{\beta} + \widehat{\alpha}^2 - 2\widehat{\alpha}\overline{y} + 2\widehat{\alpha}\overline{x}\widehat{\beta})\end{aligned}$$

$$n\bar{\hat{\epsilon}}^2 = n\left(\frac{1}{n}\sum_{k\in\mathcal{S}}(y_k - \widehat{\alpha} - x_k\widehat{\beta})\right)^2 = n\left(\overline{y} - \widehat{\alpha} - \overline{x}\widehat{\beta}\right)^2$$

So can write our variance estimator in terms of five sufficient statistics.

$$\hat{V}_{\text{greg}} = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left(n(\overline{yy} - 2\widehat{\beta}\overline{xy} + \widehat{\beta}\overline{xx}\widehat{\beta} + \widehat{\alpha}^2 - 2\widehat{\alpha}\overline{y} + 2\widehat{\alpha}\overline{x}\widehat{\beta}) - n\left(\overline{y} - \overline{x}\widehat{\beta}\right)^2\right).$$

As with the point estimate, we can privatise the variance estimate $\hat{V}_{\text{greg}}$ by privatizing the sufficient statistics and calculating a "plug-in estimator" for the variance estimate.

Sensitivity Analysis and Noise Calibration To privatize the sufficient statistics we have to provide sensitivity bound to calibrate noise for the gaussian mechanism. We will assume that data inhabits within the bounds,

$$x_k \in [-B_x, B_x], \quad y_k \in [-B_y, B_y].$$

Thus for neighboring samples $\mathcal{S} \sim \mathcal{S}$ different at individuals i and j we have,

$$\Delta_{\overline{x}} \leq |\overline{x} - \overline{x}'| = \left|\frac{1}{n}\sum_{k\in\mathcal{S}} x_k - \frac{1}{n}\sum_{k\in\mathcal{S}'} x_k\right| = \frac{|x_i - x_j|}{n} \leq \frac{2B_x}{n}$$

$$\Delta_{\overline{y}} \leq |\overline{y} - \overline{y}'| = \left|\frac{1}{n}\sum_{k\in\mathcal{S}} y_k - \frac{1}{n}\sum_{k\in\mathcal{S}'} y_k\right| = \frac{|y_i - y_j|}{n} \leq \frac{2B_y}{n}$$

$$\Delta_{\overline{xy}} \leq |\overline{xy} - \overline{xy}'| = \left|\frac{1}{n}\sum_{k\in\mathcal{S}} x_k y_k - \frac{1}{n}\sum_{k\in\mathcal{S}'} x_k y_k\right| = \frac{|x_i x_i - x_j y_j|}{n} \leq \frac{2B_x B_y}{n}$$

$$\Delta_{\overline{xx}} \leq |\overline{xx} - \overline{xx}'| = \left|\frac{1}{n}\sum_{k\in\mathcal{S}} x_k^2 - \frac{1}{n}\sum_{k\in\mathcal{S}'} x_k^2\right| = \frac{|x_i^2 - x_j^2|}{n} \leq \frac{B_x^2}{n}$$

$$\Delta_{\overline{yy}} \leq |\overline{yy} - \overline{yy}'| = \left|\frac{1}{n}\sum_{k\in\mathcal{S}} y_k^2 - \frac{1}{n}\sum_{k\in\mathcal{S}'} y_k^2\right| = \frac{|y_i^2 - y_j^2|}{n} \leq \frac{B_y^2}{n}$$

For ρ_θ-zCDP the variance of the Gaussian distribution is given by

$$\sigma_\theta^2 = \frac{\Delta_\theta^2}{2\rho_\theta}$$

. We have the following noise calibrations

$$\sigma_{\overline{x}}^2 = \frac{2B_x^2}{\rho_{\overline{x}} n^2}, \quad \sigma_{\overline{y}}^2 = \frac{2B_y^2}{\rho_{\overline{y}} n^2}, \quad \sigma_{\overline{xx}}^2 = \frac{B_x^4}{2\rho_{\overline{xx}} n^2}, \quad \sigma_{\overline{xy}}^2 = \frac{2B_x^2 B_y^2}{\rho_{\overline{xy}} n^2}, \quad \sigma_{\overline{yy}}^2 = \frac{B_y^4}{2\rho_{\overline{xx}} n^2}$$

Privacy Budgeting Given a privacy budget ε, failure allowance δ, and sensitivity bound defined for all neighboring datasets, we can translate (ε, δ)-DP to ρ-zCDP with the conversion formula $\rho = \varepsilon^2/(2\ln(1/\delta))$. Our mechanism is a composition of five different private statistics, our privacy budget has to be split up.

$$\rho = \rho_{\overline{x}} + \rho_{\overline{y}} + \rho_{\overline{xx}} + \rho_{\overline{xy}} + \rho_{\overline{yy}}.$$

In our simulation we will simplify by splitting ρ evenly over 5.

With these values we can privatize each sufficient statistics $\overline{x}, \overline{y}, \overline{xx}, \overline{xy}, \overline{yy}$ by drawing noise from their respective Normal distributions, to create $\widetilde{\overline{x}}, \widetilde{\overline{y}}, \widetilde{\overline{xx}}, \widetilde{\overline{xy}}, \widetilde{\overline{yy}}$. As the GREG estimator is a function of these sufficient statistics $\widehat{\mu}_{\text{greg}} = f(\overline{x}, \overline{y}, \overline{xx}, \overline{xy}, \overline{yy})$, we can construct a DP GREG with the "plug-in method" $\widetilde{\mu}_{\text{greg}} = f(\widetilde{\overline{x}}, \widetilde{\overline{y}}, \widetilde{\overline{xx}}, \widetilde{\overline{xy}}, \widetilde{\overline{yy}})$. We can also apply the same outline to privatize the Variance estimator for $\widehat{\mu}_{\text{greg}}$, $\hat{V}_{\text{greg}}$ to create $\tilde{V}_{\text{greg}}$, but this alone is incomplete for the true variance of the private estimator $\text{Var}(\widetilde{\mu}_{\text{greg}})$.

Privatization Algorithm Below is an outline for the algorithm discussed.

Algorithm 1 DP GREG via noisy sufficient statistics NoisyStats

---

Outputs: (ε, δ)-Differentially Private GREG estimator for SRSWOR sampling setting $\widetilde{\mu}_{\text{greg}}$

Public: Population values for the auxiliary variable x, $\{x_k\}_{k\in\mathcal{U}}$, so $\mu_x = \sum_{k\in\mathcal{U}} x_k/N$ is public. Population size $N = |\mathcal{U}|$ is also public.

Sensitive: The participants of any sample $\mathcal{S} \leftarrow \mathcal{U}$.

Require: Privacy budget ρ, ability to sample $\mathcal{U}$ for target variable y via SRSWOR.

1: Choose a sample size $n \leq N$ for SRSWOR(n), choose clipping windows for the x and y variables $[-B_x, B_x]$ and $[-B_y, B_y]$. The values n, B_x, B_y can be made public.

2: Take a SRSWOR(n) sample $\mathcal{S} \subset \mathcal{U}$ from the population. We have access to $\{(x_k, y_k)\}_{k\in\mathcal{S}}$

3: Clip the sampled values according to the choice for B_x and B_y

$$x_k := \min(\max(-B_x, x_k), B_x), \quad y_k := \min(\max(-B_y, y_k), B_y)$$

4: Split privacy budget across for the five statistics

$$\rho = \rho_{\overline{x}} + \rho_{\overline{y}} + \rho_{\overline{xx}} + \rho_{\overline{xy}} + \rho_{\overline{yy}}.$$

We will take $\rho_{\overline{x}} = \rho_{\overline{y}} = \rho_{\overline{xx}} = \rho_{\overline{xy}} = \rho_{\overline{yy}} = \rho/5$.

5: Calibrate noise by

$$\sigma^2_{\overline{x}} = \frac{2B_x^2}{\rho_{\overline{x}} n^2}, \quad \sigma^2_{\overline{y}} = \frac{2B_y^2}{\rho_{\overline{y}} n^2}, \quad \sigma^2_{\overline{xx}} = \frac{B_x^4}{2\rho_{\overline{xx}} n^2}, \quad \sigma^2_{\overline{xy}} = \frac{2B_x^2 B_y^2}{\rho_{\overline{xy}} n^2}, \quad \sigma^2_{\overline{yy}} = \frac{B_y^4}{2\rho_{\overline{xx}} n^2}$$

6: Find the true values for the following statistics

$$\overline{x} = \frac{1}{n}\sum_{k\in\mathcal{S}} x_k, \quad \overline{y} = \frac{1}{n}\sum_{k\in\mathcal{S}} y_k, \quad \overline{xy} = \frac{1}{n}\sum_{k\in\mathcal{S}} x_k y_k, \quad \overline{xx} = \frac{1}{n}\sum_{k\in\mathcal{S}} x_k^2.$$

7: Privatize each statistic by sampling noise from their respective calibrated Gaussian Distributions.

$$\begin{aligned}
\widetilde{\overline{x}} &= \overline{x} + \eta_{\overline{x}}, & \eta_{\overline{x}} &\leftarrow \mathcal{N}(0, \sigma^2_{\overline{x}}),\\
\widetilde{\overline{y}} &= \overline{y} + \eta_{\overline{y}}, & \eta_{\overline{y}} &\leftarrow \mathcal{N}(0, \sigma^2_{\overline{y}}),\\
\widetilde{\overline{xx}} &= \overline{xx} + \eta_{\overline{xx}}, & \eta_{\overline{xx}} &\leftarrow \mathcal{N}(0, \sigma^2_{\overline{xx}}),\\
\widetilde{\overline{xy}} &= \overline{xy} + \eta_{\overline{xy}}, & \eta_{\overline{xy}} &\leftarrow \mathcal{N}(0, \sigma^2_{\overline{xy}}),\\
\widetilde{\overline{yy}} &= \overline{yy} + \eta_{\overline{yy}}, & \eta_{\overline{yy}} &\leftarrow \mathcal{N}(0, \sigma^2_{\overline{yy}}).
\end{aligned}$$

8: Calculate

$$\widetilde{\beta} = \frac{\widetilde{\overline{xy}} - \widetilde{\overline{x}}(\widetilde{\overline{y}})}{\widetilde{\overline{xx}} - \widetilde{\overline{x}}^2}, \quad \widetilde{\alpha} = \overline{y} - \overline{x}\widehat{\beta}$$

$$\widetilde{\mu}_{\text{greg}} = \widetilde{\alpha} + \mu_x \widetilde{\beta}$$

$$\tilde{V}_{\text{greg}} = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\left(n(\widetilde{\overline{yy}} - 2\widetilde{\beta}\widetilde{\overline{xy}} + \widetilde{\beta}\widetilde{\overline{xx}}\widetilde{\beta} + \widetilde{\alpha}^2 - 2\widetilde{\alpha}\widetilde{\overline{y}} + 2\widetilde{\alpha}\widetilde{\overline{x}}\widetilde{\beta}) - n\left(\widetilde{\overline{y}} - \widetilde{\overline{x}}\widetilde{\beta}\right)^2\right).$$

return DP GREG Point estimate $\widetilde{\mu}_{\text{greg}}$, DP GREG variance estimate $\tilde{V}_{\text{greg}}$.

True Variance of DP GREG We previously outlined how we plan to privatize the Variance estimate for the GREG estimator. However, this alone does not describe the true variance of the DP GREG estimator which has additional variance from the noise added to the sufficient statistics. This detail can be seen in the following.

$$\begin{aligned}\mathrm{Var}(\widetilde{\mu}_{\mathrm{greg}}) &= \mathrm{Var}\left(\mathbb{E}[\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S}]\right) + \mathbb{E}\left[\mathrm{Var}(\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S})\right] \\ &\approx \mathrm{Var}\left(\widehat{\mu}_{\mathrm{greg}}\right) + \mathbb{E}\left[\mathrm{Var}(\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S})\right]\end{aligned}$$

The value $\mathrm{Var}\left(\widehat{\mu}_{\mathrm{greg}}\right)$ is estimated by $\hat{V}_{\mathrm{greg}}$, and we found a privatized version $\tilde{V}_{\mathrm{greg}}$ in our algorithm via the same "plug-in" method.

The value $\mathbb{E}\left[\mathrm{Var}(\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S})\right]$ and the slight difference between $\mathrm{Var}(\mathbb{E}[\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S}])$ and $\mathrm{Var}\left(\widehat{\mu}_{\mathrm{greg}}\right)$ remains to be discussed.

Noise contribution Start with the variance of the noise added to the privatized point estimator $\mathrm{Var}(\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S})$. This describes the net variance of the noise added to $\widehat{\mu}_{\mathrm{greg}}$ given a fixed sample $\mathcal{S}$. In our algorithm we do not add noise directly to $\widehat{\mu}_{\mathrm{greg}}$ but rather to the sufficiency statistics that compose $\widehat{\mu}_{\mathrm{greg}}$. Furthermore these statistics are composed in a nonlinear function.

$$\widehat{\mu}_{\mathrm{greg}} = f(\overline{x}, \overline{y}, \overline{xx}, \overline{xy}) = \overline{y} - (\overline{x} - \mu_x)\frac{\overline{xy} - (\overline{x})(\overline{y})}{\overline{xx} - \overline{x}^2}$$

$$\widetilde{\mu}_{\mathrm{greg}} = f(\widetilde{\overline{x}}, \widetilde{\overline{y}}, \widetilde{\overline{xx}}, \widetilde{\overline{xy}}) = \widetilde{\overline{y}} - (\widetilde{\overline{x}} - \mu_x)\frac{\widetilde{\overline{xy}} - (\widetilde{\overline{x}})(\widetilde{\overline{y}})}{\widetilde{\overline{xx}} - \widetilde{\overline{x}}^2}$$

We will use the delta method to approximate the total variance contributed by the noisy sufficient statistics.

Let

$$a = \bar{x}, \quad b = \bar{y}, \quad c = \overline{xx}, \quad d = \overline{xy}$$

and

$$f(a, b, c, d) = b - (a - \mu_x)\frac{d - ab}{c - a^2}.$$

The partial derivatives are:

$$\frac{\partial f}{\partial a} = -\frac{d - ab}{c - a^2} - (a - \mu_x)\frac{-bc - a^2 b + 2ad}{(c - a^2)^2},$$

$$\frac{\partial f}{\partial b} = 1 + \frac{a(a - \mu_x)}{c - a^2},$$

$$\frac{\partial f}{\partial c} = \frac{(a - \mu_x)(d - ab)}{(c - a^2)^2},$$

$$\frac{\partial f}{\partial d} = -\frac{a - \mu_x}{c - a^2}.$$

The gradient is:

$$\nabla f(a, b, c, d) = \begin{bmatrix} -\dfrac{d - ab}{c - a^2} - (a - \mu_x)\dfrac{-bc - a^2 b + 2ad}{(c - a^2)^2} \\ 1 + \dfrac{a(a - \mu_x)}{c - a^2} \\ \dfrac{(a - \mu_x)(d - ab)}{(c - a^2)^2} \\ -\dfrac{a - \mu_x}{c - a^2} \end{bmatrix}.$$

Finally, with independent noise variances $\sigma_a^2, \sigma_b^2, \sigma_c^2, \sigma_d^2$, the delta-method approximation for the variance is:

$$\mathrm{Var}_{\mathrm{noise}}(\tilde{\mu}_{greg}) \approx \left(\frac{\partial f}{\partial a}\right)^2 \sigma_a^2 + \left(\frac{\partial f}{\partial b}\right)^2 \sigma_b^2 + \left(\frac{\partial f}{\partial c}\right)^2 \sigma_c^2 + \left(\frac{\partial f}{\partial d}\right)^2 \sigma_d^2.$$

The delta method specifies that this approximation is calculated with the true underlying parameters $\overline{x}, \overline{y}, \overline{xx}, \overline{xy}$. However, these are sensitive quantities, so the best we can do is apply the plug in method again and use their privatized versions. $\widetilde{\overline{x}}, \widetilde{\overline{y}}, \widetilde{\overline{xx}}, \widetilde{\overline{xy}}$. The algorithm we will use to estimate the noise contribution for Variance is outlined below.

Algorithm 2 Variance contribution approximation for NoisyStats DP GREG

Outputs: An approximation for $\text{Var}(\widetilde{\mu}_{\text{greg}} \mid \mathcal{S})$, the total variance contribution to $\text{Var}(\widetilde{\mu}_{\text{greg}})$ from privatizing the sufficient statistics in the NoisyStats algorithm.

Require: Private sufficient statistics $\widetilde{\overline{x}}, \widetilde{\overline{y}}, \widetilde{\overline{xx}}, \widetilde{\overline{xy}}$, their respective noise variance calibrations $\sigma^2_{\overline{x}}, \sigma^2_{\overline{y}}, \sigma^2_{\overline{xx}}, \sigma^2_{\overline{xy}}$, and the population mean for auxiliary variable x, μ_x (public).

1: Take a size n SRSWOR sample $\mathcal{S} \subset \mathcal{U}$ from $\mathcal{U}$ for $\{y_k\}_{k \in \mathcal{S}}$.

2: Let

$$a = \widetilde{\overline{x}}, \quad b = \widetilde{\overline{y}}, \quad c = \widetilde{\overline{xx}}, \quad d = \widetilde{\overline{xy}},$$

3: Let

$$f_a = -\frac{d - ab}{c - a^2} - (a - \mu_x)\frac{-bc - a^2 b + 2ad}{(c - a^2)^2},$$

$$f_b = 1 + \frac{a(a - \mu_x)}{c - a^2},$$

$$f_c = \frac{(a - \mu_x)(d - ab)}{(c - a^2)^2},$$

$$f_d = -\frac{a - \mu_x}{c - a^2}.$$

return $\text{Var}_{\text{noise}}(\tilde{\mu}_{greg}) \approx f_a^2\sigma_a^2 + f_b^2\sigma_b^2 + f_c^2\sigma_c^2 + f_d^2\sigma_d^2$.

As a first order approximation, we can already expect our approximation to be slightly off by high order terms. The decision to use the privatized sufficient stats for the delta method approximation introduces additional complexity. Many of the gradient terms are non linear rational functions $g_a(a, b, c, d)$. As a result $\mathbb{E}[g_a(\overline{x}, \overline{y}, \overline{xx}, \overline{xy})] \neq \mathbb{E}[\widetilde{\overline{x}}, \widetilde{\overline{y}}, \widetilde{\overline{xx}}, \widetilde{\overline{xy}}]$ even though each private statistic is unbiased for their non private counterpart. The nonlinearity of the gradients come from the fractions like $1/(c - a^2)$, which are typically convex as $c - a = \overline{xx} - \overline{x}^2 \approx \text{Var}(x) > 0$. By Jensen's inequality for convex functions ϕ we expect

$$\mathbb{E}[\phi(\tilde{\theta})] > \phi(\theta).$$

Therefore, we should expect our approximation of the variance contribution to be slightly conservative or large. The final layer of complexity is that since we sample once, we will approximate $\mathbb{E}_{\mathcal{S}}[\text{Var}(\widetilde{\mu}_{\text{greg}} \mid \mathcal{S})] \approx \text{Var}(\widetilde{\mu}_{\text{greg}} \mid \mathcal{S})$ for n large enough. How these fine points interact overall remains to be observed in simulation.

Slight Bias in the Point Estimate The other slight point in the Variance decomposition is a slight difference between $\text{Var}(\mathbb{E}[\widetilde{\mu}_{\text{greg}} \mid \mathcal{S}])$ and $\text{Var}\left(\widehat{\mu}_{\text{greg}}\right)$. We can estimate $\text{Var}\left(\widehat{\mu}_{\text{greg}}\right)$ with $\hat{V}_{\text{greg}}$ which we can privatize as $\tilde{V}_{\text{greg}}$. However, we might expect some bias in the point estimate i.e. $\mathbb{E}[\widetilde{\mu}_{\text{greg}}|\mathcal{S}] \neq \widehat{\mu}_{\text{greg}}|\mathcal{S}$. This is because $\widehat{\mu}_{\text{greg}}$ is a nonlinear function of the sufficient statistics which we are privatizing. To estimate this bias (and the variance of that bias) we could use the delta method again, but we can expect that our estimated bias to also be a nonlinear function of the true sufficient statistics. Privatizing any correction for bias would also result in a slight bias. We ran in to the same issue for approximating $\text{Var}(\widetilde{\mu}_{\text{greg}} \mid \mathcal{S})$ after the delta method, and chose not to correct for this nonlinear-expectation bias. We take the same approach here: we do not attempt to correct the slight bias in the point estimate and instead assess performance empirically through simulation.

3.2 Algorithm 2: NoisyModel

Rather than injecting noise for the sub-components of the model, this second algorithm seeks to directly inject noise to $\widehat{\mu}_{\text{greg}}$. This requires a bound on the sensitivity for $\widehat{\beta}$, and turns out to be difficult to do theoretically. To do this, we will borrow a result from cross validation, the leave one out estimator.

Theorem 3.1 (The leave one out estimator) *Let $\mathcal{S}$ be a n sized sample from $\mathcal{U}$, and let $\widehat{\beta}$ be the OLS estimator solution of*

$$\widehat{\beta} = \underset{\mathbf{u} \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{k \in \mathcal{S}} (y_k - \mathbf{x}_k^\top \mathbf{u})^2.$$

We also define $\widehat{\beta}_0$ as the OLS estimator solution for $\mathcal{S}_0 = \mathcal{S} \backslash \{i\}$

$$\widehat{\beta}_0 = \underset{\mathbf{u} \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{k \in \mathcal{S}_0} (y_k - \mathbf{x}_k^\top \mathbf{u})^2.$$

We can then get the following result

$$\widehat{\beta}_0 = \widehat{\beta} - \frac{(X^\top X)\mathbf{x}_k}{1 - h_{ii}} \widehat{\epsilon}_i, \quad h_{ii} = \mathbf{x}_i (X^\top X)^{-1} \mathbf{x}_i, \quad \widehat{\epsilon}_i = y_i - \mathbf{x}_i^\top \widehat{\beta}.$$

In our setting where we also have a neighboring dataset $\mathcal{S}' = \mathcal{S}_0 \cup \{j\}$ and its corresponding model $\widehat{\beta}'$, we also have

$$\widehat{\beta}_0 = \widehat{\beta}' - \frac{(X'^\top X')^{-1}\mathbf{x}_k}{1-h'_{jj}}\widehat{\epsilon}'_j, \quad h'_{jj} = \mathbf{x}_j(X'^\top X')^{-1}\mathbf{x}_j, \quad \widehat{\epsilon}'_j = y_j - \mathbf{x}_j^\top\widehat{\beta}'.$$

$$\|\widehat{\beta} - \widehat{\beta}'\| = \left\|\frac{(X^\top X)^{-1}\mathbf{x}_k}{1-h_{ii}}\widehat{\epsilon}_i - \frac{(X'^\top X')^{-1}\mathbf{x}_k}{1-h'_{jj}}\widehat{\epsilon}'_j\right\| = \left\|\frac{A^{-1}\mathbf{x}_k}{1-h_{ii}}\widehat{\epsilon}_i - \frac{A'^{-1}\mathbf{x}_k}{1-h'_{jj}}\widehat{\epsilon}'_j\right\| \leq \left\|\frac{A^{-1}\mathbf{x}_k}{1-h_{ii}}\widehat{\epsilon}_i\right\| + \left\|\frac{A'^{-1}\mathbf{x}_k}{1-h'_{jj}}\widehat{\epsilon}'_j\right\|$$

Private Point Estimate Construction Unlike the first algorithm, we will include the intercept within the OLS coefficient in this model. So $\mathbf{x}_k = [1, a_k]^\top$, and any bound on $\mathbf{x}_k$ account for this, $B_\mathbf{x} = \sqrt{1+B_a^2}$. With the necessary bounds, we can calibrate the noise required to privatize the GREG statistic.

We an give the spectral bound on $\|(X^\top X)\| = \|A\|$ with its minimum possible eigenvalue $\lambda_{\min}$. Similarly, we can bound $\|A^{-1}\|$ by $\lambda_{\min}$. We can provide the value $1/\lambda_{\min}$ in our setting because the $\mathbf{x}_k$'s are assumed to be public, so the attacker can find this value too. We have

$$h_{k\ell} = \mathbf{x}_k A^{-1}\mathbf{x}_\ell \leq B_\mathbf{x}\frac{1}{\lambda_{\min}}B_\mathbf{x} = \frac{B_\mathbf{x}}{\lambda_{\min}} =: h_{\max}.$$

If we bound $\widehat{\epsilon}_k$ by $2B_y$, we get the sensitivity bound,

$$\Delta_{\widehat{\beta}} = \max_{\mathcal{S}\sim\mathcal{S}'\leftarrow\mathcal{P}(\mathcal{U})}\|\widehat{\beta}-\widehat{\beta}'\| \leq 2\frac{1}{\lambda_{\min}}B_\mathbf{x}2B_y\frac{1}{1-h_{\max}} = \frac{4B_\mathbf{x}B_y}{\lambda_{\min}-\lambda_{\min}\frac{B_\mathbf{x}^2}{\lambda_{\min}}} = \frac{4B_\mathbf{x}B_y}{\lambda_{\min}-B_\mathbf{x}^2}.$$

In the setting where the intercept is included in the OLD model, we can write the greg estimator as

$$\widehat{\mu}_{\text{greg}} = \mu_x^\top\widehat{\beta}$$

So given $\lambda_{\min}$, we can give at differentially private $\widehat{\mu}_{\text{greg}}$ by directly injecting noise to $\widehat{\beta}$.

Private Variance Estimate Construction Using the same method, we will seek to add noise directly to the Variance estimate $\widehat{V}_{\text{greg}} = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{k\in\mathcal{S}}\left(\widehat{\epsilon}_k - \overline{\widehat{\epsilon}}\right)^2$ Since the intercept is included in the OLS coefficient in this set-up we can use the fact that $\overline{\widehat{\epsilon}} = 0$. The Variance estimate we need to do sensitivity analysis on is

$$\widehat{V}_{\text{greg}} = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{k\in\mathcal{S}}\widehat{\epsilon}_k^2.$$

For neighboring samples $\mathcal{S}$ and $\mathcal{S}'$, we can provide on $\widehat{\epsilon}_k^2 - \widehat{\epsilon}_k'^2$.

$$\begin{aligned}
|\widehat{\epsilon}_k^2 - \widehat{\epsilon}_k'^2| &= |(y_k - \mathbf{x}_k^\top\widehat{\beta})^2 - (y_k - \mathbf{x}_k^\top\widehat{\beta}')^2| \\
&= |y_k^2 - 2y_k\mathbf{x}_k^\top\widehat{\beta} + (\mathbf{x}_k^\top\widehat{\beta})^2 - y_k^2 + 2y_k\mathbf{x}_k^\top\widehat{\beta}' - (\mathbf{x}_k^\top\widehat{\beta}')^2| \\
&\leq 2y_k\mathbf{x}_k\|\widehat{\beta}' - \widehat{\beta})\| + |(\mathbf{x}_k^\top\widehat{\beta})^2 - (\mathbf{x}_k^\top\widehat{\beta}')^2| \\
&\leq 2B_yB_\mathbf{x}\frac{4B_\mathbf{x}B_y}{\lambda_{\min}-B_\mathbf{x}^2} + \left(B_\mathbf{x}\frac{4B_\mathbf{x}B_y}{\lambda_{\min}-B_\mathbf{x}^2}\right)^2
\end{aligned}$$

$$\begin{aligned}
\Delta_{\widehat{V}} &= |\widehat{V} - \widehat{V}'| \\
&= \left|\frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left[\sum_{k\in\mathcal{S}\cap\mathcal{S}'}\left(\hat{\epsilon}_k^2 - \hat{\epsilon}_k'^2\right) + \hat{\epsilon}_j^2 - \hat{\epsilon}_i'^2\right]\right| \\
&\leq \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left[\sum_{k\in\mathcal{S}\cap\mathcal{S}'}\left|\frac{B_x^2B_e}{\lambda_{\min}-B_x^2}\right|^2 + \hat{\epsilon}_j^2 - \hat{\epsilon}_i'^2\right] \\
&\leq \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left[(n-1)\left(\frac{8B_\mathbf{x}^2B_y^2}{\lambda_{\min}-B_\mathbf{x}^2} + \left(\frac{4B_\mathbf{x}^2B_y}{\lambda_{\min}-B_\mathbf{x}^2}\right)^2\right) + (2B_y)^2\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left(\frac{8B_\mathbf{x}^2B_y^2}{\lambda_{\min}-B_\mathbf{x}^2} + \left(\frac{4B_\mathbf{x}^2B_y}{\lambda_{\min}-B_\mathbf{x}^2}\right)^2\right) + \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}4B_y^2.
\end{aligned}$$

Variance estimate correction As with algorithm 1, since we added noise to the point estimate in order to privatize it, the variance of that noise should be reflected in the variance estimate of the DP GREG estimator. The construction of each point and variance estimate to have mean 0 noise added at the end means that out estimates for each are unbiased, this pays dividends in the following decomposition.

$$\begin{aligned}
\mathrm{Var}(\widetilde{\mu}_{\mathrm{greg}}) &= \mathrm{Var}\left(\mathbb{E}[\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S}]\right) + \mathbb{E}\left[\mathrm{Var}(\widetilde{\mu}_{\mathrm{greg}} \mid \mathcal{S})\right] \\
&= \mathrm{Var}\left(\mathbb{E}[\mu_{\mathbf{x}}^\top(\widehat{\beta} + \eta_{\widehat{\beta}}) \mid \mathcal{S}]\right) + \mathbb{E}\left[\mathrm{Var}(\mu_{\mathbf{x}}^\top(\widehat{\beta} + \eta_{\widehat{\beta}}) \mid \mathcal{S})\right] \\
&= \mathrm{Var}\left(\mathbb{E}[\mu_{\mathbf{x}}^\top\widehat{\beta} \mid \mathcal{S}] + \mathbb{E}[\mu_{\mathbf{x}}^\top\eta_{\widehat{\beta}} \mid \mathcal{S}]\right) + \mathbb{E}\left[\mu_{\mathbf{x}}^2\,\mathrm{Var}(\eta_{\widehat{\beta}} \mid \mathcal{S})\right] \\
&= \mathrm{Var}\left(\mathbb{E}[\widehat{\mu}_{\mathrm{greg}} \mid \mathcal{S}] + 0\right) + \mathbb{E}\left[\mu_{\mathbf{x}}^2\sigma_{\widehat{\beta}}^2\right] \\
&= \mathrm{Var}\left(\widehat{\mu}_{\mathrm{greg}}\right) + \mu_{\mathbf{x}}^2\sigma_{\widehat{\beta}}^2 \\
&\approx \widehat{V} + \mu_{\mathbf{x}}^2\sigma_{\widehat{\beta}}^2 \quad \text{(By the asymptotically unbiased estimate } \widehat{V}\text{)}
\end{aligned}$$

This is a much cleaner variance estimate than that for the previous algorithm with not only has a slightly biased point estimate that was to accounted for but also an estimated value for the second term via the delta method. Our final private variance estimate for the DP GREG estimator is thus

$$\widehat{\mathrm{Var}}(\widetilde{\mu}_{\mathrm{greg}}) = \widetilde{V} + \mu_x^2\sigma_{\widehat{\beta}}^2$$

Algorithm 3 DP GREG Estimator via NoisyModel

Public: $\{\mathbf{x}_k\}_{k\in\mathcal{U}}$, and sample size n, bounds for $|y_k|$, $\|\mathbf{x}_k\|$, over $k \in \mathcal{U}$ as B_y, and $B_{\mathbf{x}}$ respectively. A bound on the augmented auxiliary variable matrix $X^\top X$, $\lambda_{\min}$.

Require: Privacy budget ϵ, failure allowance δ, access to sample $\{y_k\}_{k\in\mathcal{S}}$ for $\mathcal{S} \subset \mathcal{U}$.

Ensure: Approximate Differentially Private GREG estimate $\widetilde{\mu}_{\mathrm{greg}}$.

1: Take a size n SRSWOR sample $\mathcal{S} \subset \mathcal{U}$ from $\mathcal{U}$ for $\{y_k\}_{k\in\mathcal{S}}$.

2: Calculate the model statistics

$$\widehat{\beta} = \left(\sum_{k\in\mathcal{S}} \mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\left(\sum_{k\in\mathcal{S}} \mathbf{x}_k y_k\right)$$

$$\widehat{V}_{\mathrm{greg}} = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_{k\in\mathcal{S}}(y_x - \mathbf{x}_k^\top\widehat{\beta})^2$$

3: Convert to ρ-zCDP by

$$\rho = \varepsilon^2/(2\ln(1/\delta))$$

4: Split the privacy budge for the two statistics. $\rho = \rho_{\widehat{\beta}} + \rho_{\widehat{V}} = 2\rho_{\widehat{\beta}}$

5: Estimate sensitivity for $\widehat{\beta}$ and $\widehat{V}$,

$$\Delta_{\widehat{\beta}} \le \frac{2B_{\mathbf{x}}B_\epsilon}{\lambda_{\min} - B_{\mathbf{x}}^2}$$

$$\Delta_{\widehat{V}} \le \frac{1}{n}\left(1 - \frac{n}{N}\right)\left(\frac{8B_{\mathbf{x}}^2B_y^2}{\lambda_{\min} - B_{\mathbf{x}}^2} + \left(\frac{4B_{\mathbf{x}}^2B_y}{\lambda_{\min} - B_{\mathbf{x}}^2}\right)^2\right) + \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}4B_y^2$$

6: Compute noise variance

$$\sigma_{\widehat{\beta}}^2 := \frac{\Delta_{\widehat{\beta}}^2}{2\rho}, \quad \sigma_{\widehat{V}}^2 := \frac{\Delta_{\widehat{V}}^2}{2\rho}$$

7: Sample noises $\eta_{\widehat{\beta}} \leftarrow \mathcal{N}(0, \sigma_{\widehat{\beta}}^2 I)$, and $\eta_{\widehat{V}} \leftarrow \mathcal{N}(0, \sigma_{\widehat{V}}^2 I)$

8: Set $\widetilde{\beta} := \widehat{\beta} + \eta_{\widehat{\beta}}$ and $\widetilde{V} := \widehat{V} + \eta_{\widehat{V}}$

9: Compute $\widetilde{\mu}_{\mathrm{greg}} := \mu_x^\top\widetilde{\beta}$, and $\widehat{\mathrm{Var}}(\widetilde{\mu}_{\mathrm{greg}}) = \widetilde{V} + \mu_x^2\sigma_{\widehat{\beta}}^2$

10: **return** $\widetilde{\mu}_{\mathrm{greg}}$, and $\widehat{\mathrm{Var}}(\widetilde{\mu}_{\mathrm{greg}})$

Theoretic advantages and challenges Structurally, this second algorithm, NoisyModel, has a cleaner variance and point estimate. They are not only unbiased but also not mere first order estimations. Moreover, this algorithm has a simple extension to higher dimensions. (NoisyStats in higher dimensions would additionally require privatizing the matrix $\overline{\mathbf{xx}} = \frac{1}{n}\sum_{k\in\mathcal{S}} \mathbf{x}_k\mathbf{x}_k^\top$.) The main theoretic drawback of NoisyModel is with regards to $\lambda_{\min}$. We state

that this value can be made public because $\{\mathbf{x}_k\}_{k\in\mathcal{U}}$, and n are public. However, this value can be difficult to find in practice for large populations and countless possible samples. In our simulations we will first run an empirical simulation to find a values for $\lambda_{\min}$. This does, unfortunately, deviate from the spirit of DP and its emphasis on absolute worst-case-bounds. What we get instead is an approximate DP where our privacy guarantee holds expect in very rare and unfortunate samples.

3.3 Algorithm 3: NoisySample

This last algorithm simply publicizes a privatized synthetic sample. Calculations of the point estimate and confidence intervals proceed without adjustments as the sample itself is already injected with Gaussian noise.

Sensitivity for $\mathcal{S}$ Our algorithm proposes that we privatize the sample $\{(x_k, y_k)\}_{k\in\mathcal{S}}$. We do this by privatizing the vectors $\mathbf{x}_S = (x_1, x_2, ..., x_n)^\top$ and $\mathbf{y}_S = (y_1, y_2, ..., y_n)^\top$ with the following sensitivities,

$$\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}_{\mathcal{S}'}\| = \|x_i - x_j\| \leq 2B_x$$

$$\|\mathbf{y}_{\mathcal{S}} - \mathbf{y}_{\mathcal{S}'}\| = \|y_i - y_j\| \leq 2B_y$$

The algorithm for the GREG estimator confidence intervals is as follows,

Algorithm 4 DP GREG Estimator via NoisySample

Public: $\{x_k\}_{k\in\mathcal{U}}$, and sample size n, bounds for $|y_k|$, $\|x_k\|$, over $k \in \mathcal{U}$ as B_y, and B_x respectively.
Require: Privacy budget ρ, access to sample $\{y_k\}_{k\in\mathcal{S}}$ for $\mathcal{S} \subset \mathcal{U}$.
Ensure: Approximate Differentially Private GREG estimate $\widetilde{\mu}_{\text{greg}}$.

1: Split the privacy budge for the two statistics. $\rho = \rho_x + \rho_y = 2\rho_x$
2: Take a size n SRSWOR sample $\mathcal{S} \subset \mathcal{U}$ from $\mathcal{U}$. $\mathbf{x} = (x_1, ..., x_n)$, $\mathbf{y} = (y_1, ..., y_n)$.
3: Let

$$\Delta_x = 2B_x, \quad \Delta_y = 2B_y.$$

4: Sample noise from Normal distribution according to $\vec{\eta}_x \leftarrow \mathcal{N}_n\left(0, \frac{\Delta_x}{2\rho_x}\right)$, $\quad \vec{\eta}_y \leftarrow \mathcal{N}_n\left(0, \frac{\Delta_y}{2\rho_y}\right)$.
5: Private the sample by

$$(\tilde{x}_1, ..., \tilde{x}_n) = (x_1, ..., x_n) + \vec{\eta}_x, \quad (\tilde{y}_1, ..., \tilde{y}_n) = (y_1, ..., y_n) + \vec{\eta}_y$$

6: Calculate the model statistics

$$\widetilde{\beta} = \left(\sum_{k\in\mathcal{S}} \tilde{x}_k \tilde{x}_k^\top\right)^{-1} \left(\sum_{k\in\mathcal{S}} \tilde{x}_k \tilde{y}_k\right)$$

$$\widetilde{V}_{\text{greg}} = \frac{1}{n}\left(1 - \frac{n}{N}\right) \frac{1}{n-1} \sum_{k\in\mathcal{S}} (y_x - \mathbf{x}_k^\top \widetilde{\beta})^2$$

7: Compute $\widetilde{\mu}_{\text{greg}} := \mu_x^\top \widetilde{\beta}$, and $\widehat{\text{Var}}(\widetilde{\mu}_{\text{greg}}) = \widetilde{V}$
8: **return** $\widetilde{\mu}_{\text{greg}}$, and $\widehat{\text{Var}}(\widetilde{\mu}_{\text{greg}})$

4 Experimental Sample Level DP GREG

This section details the implementation and results of the outlined algorithms in R simulation. We discuss preliminary observations and metrics of the Monte Carlo runs.

4.1 Experimental Setting

The three algorithms for DP GREG were compared on simulated populations of size $N = 10000$. The SRSWOR sample sizes are fixed to $n = 500$, and a total of 10000 simulations where ran.

Population Generation We construct a finite population $U = \{1, \ldots, N\}$ by first generating auxiliaries x_k under one of three regimes. For `normal`, we draw $x_k \leftarrow \mathcal{N}(0,\, 0.44^2)$; for `uniform`, we draw $x_k \leftarrow \text{Unif}[-B_x, B_x]$; and for `exponential`, we draw $x_k^\star \leftarrow \text{Exp}(1)$ and recentre via $x_k = x_k^\star - \overline{x^\star}$ to obtain mean ≈ 0 prior to clipping. Outcomes follow a linear model

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k, \qquad \varepsilon_k \leftarrow \mathcal{N}(0, \sigma_e),$$

so that the non-private superpopulation relation is controlled by (β_0, β_1) (e.g., -0.72 intercept, 2.1 slope in our default runs). The auxiliary frame $\{x_k\}_{k\in U}$ is treated as public, consistent with the GREG set-up where $\mu_x = \frac{1}{N}\sum_{k\in U} x_k$ is known.

In our simulation runs below we had $N = 10000$, $B_x = 1$, $\sigma_\epsilon = 0.44$, $\beta_0 = -1.44$, $\beta_1 = 0.42$, and $B_y = 3$

Clipping To enforce global sensitivity bounds used for DP calibration, we clip x_k to $[-B_x, B_x]$ (e.g., $B_x = 1$) and y_k to $[-B_y, B_y]$ (e.g., $B_y = 3$). Formally,

$$x_k \leftarrow \min\{\max(x_k, -B_x),\, B_x\}, \qquad y_k \leftarrow \min\{\max(y_k, -B_y),\, B_y\}.$$

Crucially, the choice of B_y is made *independently of any sampled* y_k to avoid privacy leakage; in practice it should be specified a priori from domain knowledge or policy. While too-small B_y can distort the underlying linear relation and bias estimation, setting B_y overly large inflates the noise scale in the Gaussian mechanism. Our defaults balance these considerations by keeping clipping infrequent (given the normal error with variance σ_e) while providing finite, public bounds for sensitivity.

Privacy We parameterize privacy using ρ-zCDP and report (ε, δ) via the standard conversion $\varepsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$. In our simulations we fix $\rho = 0.04342945$, which corresponds to $(\varepsilon, \delta) \approx (1.458,\, 10^{-5})$-DP. This as a slightly conservative setting for a target budget of $(\varepsilon = 1.5, \delta = 10^{-5})$.

Minimum Eigenvalue estimation for NoisyModel (A2) For algorithm 2 Noisy Model, we stated that we need the value $\lambda_{\min}$. This value is difficult to calculate, since it would requires finding the minimum eigenvalue of the matrix of $\mathbf{x}_k$'s over all possible samples. A theoretical worst case bound could be supposing the n individuals with the smallest ex_k's were selected. If there are 500 or so individuals in the population with $\mathbf{x}_k$'s very close to 0, our resulting bound would be minuscule, and the added noise would have to be enormous. We opted to seek a empirical approximation for $\lambda_{\min}$. This involves a Monte-Carlo process to repeatedly taking samples of $\{\mathbf{x}_k\}_{k\in\mathcal{U}}$ and finding the minimum eigenvalues. This approach, however, is alarming form a DP standpoint. DP has an emphasis on worst case bounds. An empirical estimation for $\lambda_{\min}$ that hold for 10000 simulations does not suffice. One possible solution could be privately releasing the $\lambda_{\min}$ found in our sample by some Gaussian mechanism. Beside having to allocate more of our privacy budge away, this would also make our sensitivities for $\widehat{\beta}$, and $\widehat{V}$ randomized which also may disqualify this algorithm from DP. Again, for our initial exploration, our simulations configure $\lambda_{\min}$ through a Monte-Carlo estimation. For the uniformly generated data, we had $\lambda_{\min} = 140$, for normally generated data we had $\lambda_{\min} = 72$, and for exponentially generated data $\lambda_{\min} = 175$.

4.2 Simulation result metrics

The results of the simulations are detailed below, with the following metrics recorded for each estimator:

- **Bias** – mean deviation of the estimator from the true mean.
- **RMSE** – root mean squared error, reflecting combined bias and variance.
- **Variance** – empirical variance of the estimator across simulation runs.
- **Coverage** – proportion of confidence intervals containing the true mean (nominal level 95%).
- **Mean CI Width** – average width of the reported 95% confidence intervals.
- **Relative Efficiency** – ratio of the DP estimator's variance to the non-private baseline variance (Algorithm 1's non-private run for the same distribution).

Additional **variance metrics** were also computed:

- **Var Bias** – difference between the reported and empirical variance.
- **Var Var** – variance of the reported variance estimates across simulation runs.
- **Empirical Var** – variance computed directly from the simulated estimates.
- **Mean Reported Var** – average variance value reported by the estimator.

The tables below show results for uniformly generated x_k values, followed by the same metrics for normally distributed and centred exponentially distributed x_k populations. The latter two exhibit similar trends to the uniform case, verifying that the estimators can capture the mean under different data-generating settings.

Table 1: Simulation results (Uniform x_k). Metrics compare Non-Private baseline (A0) and DP algorithms (A1–A3).

Metric	**A0: NP**	**A1: NoisySuffStats**	**A2: NoisyModel**	**A3: NoisySample**
Point Estimator Performance				
True mean	−1.435	−1.435	−1.435	−1.435
Mean estimate	−1.44	−1.44	−1.44	−1.44
% Relative bias	−0.018%	−0.308%	−0.564%	−0.323%
Empirical variance	0.000371	0.00870	0.326	1.58
Relative efficiency (var. ratio DP/NP)	1.0	23.5	871	4173
MSE	0.000371	0.00872	0.326	1.58
Variance Estimator Performance				
Mean variance estimate	0.000366	0.00878	0.327	1.52
Calibration κ (mean V / emp. var)	0.988	1.01	1.00	0.960
Confidence Interval Performance				
Coverage (95% nominal)	94.8%	9.52%	95.1%	94.5%
Average CI width	0.0759	0.367	2.24	4.83

Table 2: Simulation results (Normal x_k). Metrics compare Non-Private baseline (A0) and DP algorithms (A1–A3).

Metric	**A0: NP**	**A1: NoisySuffStats**	**A2: NoisyModel**	**A3: NoisySample**
Point Estimator Performance				
True mean	−1.440	−1.440	−1.440	−1.440
Mean estimate	−1.44	−1.45	−1.46	−1.44
% Relative bias	−0.00938%	−0.520%	−1.58%	0.739%
Empirical variance	0.000361	0.00911	1.25	1.56
Relative efficiency (var. ratio DP/NP)	1.0	25.2	3475	4131
MSE	0.000361	0.00916	1.25	1.56
Variance Estimator Performance				
Mean variance estimate	0.000366	0.00932	1.23	1.52
Calibration κ (mean V / emp. var)	1.01	1.02	0.990	0.971
Confidence Interval Performance				
Coverage (95% nominal)	95.2%	95.5%	94.6%	95%
Average CI width	0.0750	0.378	4.36	4.83

Table 3: Simulation results (Exponential x_k). Metrics compare Non-Private baseline (A0) and DP algorithms (A1–A3).

Metric	**A0: NP**	**A1: NoisySuffStats**	**A2: NoisyModel**	**A3: NoisySample**
Point Estimator Performance				
True mean	−1.500	−1.500	−1.500	−1.500
Mean estimate	−1.50	−1.50	−1.50	−1.49
% Relative bias	−0.0152%	−0.252%	−0.452%	0.657%
Empirical variance	0.000362	0.00863	0.217	1.56
Relative efficiency (var. ratio DP/NP)	1.0	23.8	600	2927
MSE	0.000362	0.00865	0.217	1.56

Metric	A0: NP	A1: NoisySuffStats	A2: NoisyModel	A3: NoisySample
Variance Estimator Performance				
Mean variance estimate	0.000362	0.00868	0.213	1.52
Calibration κ (mean V / emp. var)	1.00	1.01	0.984	0.970
Confidence Interval Performance				
Coverage (95% nominal)	95.1%	95.1%	94.7%	94.9
Average CI width	0.0746	0.366	1.81	4.83

4.3 Observations

Based on these preliminary observations, we can confidently discount NoisyModel (A2) and NoisySample (A3) as unworkable methods in their current form. NoisyStats (A1), despite being significantly less efficient that the non private model (which we expect), remains potentially usable. This conclusion comes down to observations regarding each models performance with regards to

Bias: Our algorithms all exhibited low Relative Bias. We know that (A2), at least is theoretically unbiased. (A1) we recall is theoretically slightly biased due to the nonlinear composition of the sufficient statistics in the model, and (A3) also has the nonlinear composition of sufficient statistics on a private sample which may not reflect the true sample. Despite this, (A1) and (A2) have empirically low biases, and performed on-par with (A2) and often better. Whatever biases existed for (A1) or (A3) did not significantly register in this particular experimental setting. Perhaps we might see the bias with more lopsided distributions of x_k.

Variance: Variance made up most of the MSE all the algorithms. While each method does a good job in estimating its own variance for confidence intervals, their variances differ significantly. Non private (A0) is much better than (A1) which is better than (A2) which is better than (A3). NoisyStats (A1) seems to benefit from having a robust theory, privatizing only five to six simple statistics and composing them. NoisyModel (A2) seems to be particularly impacted by different $\lambda_{\min}$ which impact how much noise was added to the signal. The empirically estimated $\lambda_{\min}$ was smallest for Normal x_k and that is where (A2) did the worst. NoisySample (A3) is particularly poor as a result of the simple privatization of the whole sample causing much of the variance.

Efficiency: Comparing the variances of the private models to the non private model shows how much we lose from privatization. In this metric, (A2) and (A3) are abysmal often hundreds or even thousands of times less efficient than the non private estimator. More modestly, (A1) had around $20-30$ times more variance than the non private (A0). It is important to note that Efficiency scores can vary wildly also depending on the experimental setting. In our run, the clipping window kept the vast majority of the data. If we set $\sigma_\epsilon = 1$, we get (A1) to only have 6 times more variance than (A0). This improvement is due to (A0) also performing much worse in the less linear setting.

Confidence Intervals: All of methods exhibited properly calibrated confidence intervals. They differ significantly in terms of confidence interval length. In our setting, the window clipping for y_k is $[-3; 3]$, for a window width of 6. The Non private estimator has interval widths of around 0.075, while (20 times less efficient) NoisyStats had lengths around 0.370 which is about 6% of the domain for a sample size of 500. The other two, (A2) and (A3), show the consequences of poor variance performs here with lengths of often more than a third of the domain.

5 Theoretical Population Level DP GREG

In our construction of sample level DP GREG estimation, the only item we did not need to privatize was $\mu_{\mathbf{x}}$ the population mean. This was not only available to us to provide model assisted estimation, but also public so we did not privatize this quantity. We may want to entertain the idea of providing a population level DP guarantee for the GREG estimator especially if we also wanted to protect $\{\mathbf{x}_k\}_{k\in\mathcal{U}}$.

In the case of bounded population level DP (i.e. population size N is fixed and possibly public) we can easily implement this for the NoisyStats algorithm by simply privatizing μ_x

Bounded populations are useful for sensitivity analysis of the mean,

$$\Delta_{\mu_x} = \max_{\mathcal{U} \sim \mathcal{U}'} \left| \frac{1}{N} \sum_{k \in \mathcal{U}} x_k - \frac{1}{N} \sum_{k \in \mathcal{U}'} x_k \right|$$

$$\Delta_{\mu_x} \leq 2 \left| \frac{B_x}{N} \right| = \frac{2B_x}{N}$$

Algorithm 5 Private auxiliary variable population mean for bounded population level DP f NoisyStats

Public: Bound on x_k, B_x, population size N

Require: Privacy budget ρ_{μ_x}, $\{x_k\}_{k \in \mathcal{U}}$

Ensure: Approximate Differentially Private mean μ_x at the population level for fixed sized populations $\mathcal{U}$

1: Let

$$\Delta_{\mu_x} = \frac{2B_x}{N}.$$

2: Sample noise from Normal distribution according to

$$\eta_{\mu_x} \leftarrow \mathcal{N}\left(0, \frac{\Delta_{\mu_x}^2}{2\rho_{\mu_x}}\right).$$

3: Private the signal by

$$\widetilde{\mu}_x = \mu_x + \eta_{\mu_x}$$

4: **return** $\widetilde{\mu}_x$

Replacing μ_x for $\widetilde{\mu}_x$ should give us a population level DP GREG estimate. This privatization does, however, also increase the variance. Additionally, for each algorithm, the variance estimator should also be corrected to reflect the impact of adding noise to μ_x. In the spirit of the NoisyStats algorithm, this would require another delta-method approach.

The ethos of the NoisyModel algorithm might be adapted to incorporate bounding μ_x within the sensitivity analysis of $\widehat{\mu}_{\text{greg}}$. However, an issue that may render serious reworking of Algorithm 2, that $\lambda_{\min}$ is not longer available to the public. An even stricter theoretic bound may be required. On the other hand, an adapted NoisySample algorithm would inspire adding noisy to every x_k in the population.

Despite the increased complexity and variance as a result of having to privatize μ_x, we can receive dividends from the privacy amplification by sub-sampling theorem 2.5

Since neighboring populations $\mathcal{U}$,$\mathcal{U}'$ are of the sample size, respective sample $\mathcal{S} \subset \mathcal{U}$, and $\mathcal{S}' \subset \mathcal{U}'$ exist where $\mathcal{S}$ and $\mathcal{S}'$ are also neighboring. The second assumption of the theorem is also satisfied since replacing one individual of the population does not affect the probability that other individuals are sampled; $\pi_k = n/N$ always.

As a result of this, our population level DP GREG enjoys more privacy than originally configured.

6 Discussion

6.1 Project challenges

The biggest challenge is developing a method for DP linear regression. The model is a complex function of the dataset that can be difficult to bound. Apart from NoisyStats there are other previous work on DP linear regression that is worth looking into.

I also found incorporating the privacy amplification by subsampling difficult to fully grasp. This is worth looking into especially when considering privacy guarantees at higher population levels.

I also worked on a very narrow subproblem of the larger topic of model assisted estimators. Other sampling designs will change the variance calculations, and we may also have to consider the π_k's in our bounds. For data dependent sampling methods like IPPS, there is the added issue that the sampling design leaks information about the population.

Increasing the dimension also offers challenges, the bound on the auxiliary variables is a norm bound which may mean more noise being added. A even less useful noisier signal may be the result.

As a next step, I would construct different DP GREG models for srswor based on the alternative methods in the literature. It would be worth testing the algorithms on realistic non synthetic datasets like income vs age. This would pave the way for increasing dimensions and generalizing for other sampling designs.

For a more applied result, there are other privacy guarantees used in practice that have more relaxed prerequisites like *smooth sensitivity* which may help with utility.

I have also yet to actually implement the population level DP proposal. It might be worth seeing how that performs and if it would be improved by a different DP design.

6.2 Useful Reading

A good starting point is Yu-Xiang Wang [10] for DP OLS estimator. This source has a brief review of the prior methods for DP OLS with links to their original papers. This included the Noisy-Sufficient-Stats method and noisy stochastic gradient descent among other methods. They then work on "adaptive" methods which work in unbounded domains. There is some work I do not understand about releasing a private $\lambda_{\min}$. I am not sure if the adaptive DP approach they take offers the same privacy guarantee as regular DP. Still, this work is worth looking into as a survey of DP OLS.

The paper by Alabi, ..., Vadhan [1] was where i first found the Noisy Sufficient Stats method.

DP in the sampling setting has mostly been worked on by Drechsler and Ballie. Their working paper "Whose data is it anyway" [3] introduces a clean notational rigor. They also discuss their ideas in [6].

Shrong's paper was very influential as well [8]. They had a different problem setting though. They estimated proportions. A model assisted estimator in that setting may have more of a logistic classification model instead.

The sources can be organized by: **DP background:** [9], [7], [4]; **DP in Sampling:** [3],[5],[6],[8]; **DP Linear Regression:**[1],[2],[10]

References

[1] Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially private simple linear regression. *arXiv preprint arXiv:2007.05157*, 2020.

[2] Daniel G. Alabi and Salil P. Vadhan. Differentially private hypothesis testing for linear regression. *Journal of Machine Learning Research*, 24:1–50, 2023.

[3] James Bailie and Jörg Drechsler. Whose data is it anyway? towards a formal treatment of differential privacy for surveys. Working paper, May 13, 2024.

[4] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, NeurIPS 2018, 2018.

[5] Jörg Drechsler. Differential privacy for government agencies—are we there yet? *Journal of the American Statistical Association*, 118(541):761–773, 2023.

[6] Jörg Drechsler and James Bailie. The complexities of differential privacy for survey data. NBER Working Paper 32905, National Bureau of Economic Research, sep 2024.

[7] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[8] Shurong Lin, Mark Bun, Marco Gaboardi, Eric D. Kolaczyk, and Adam Smith. Differentially private confidence intervals for proportions under stratified random sampling. *Electronic Journal of Statistics*, 18(1):1455–1494, 2024. Also available as arXiv:2301.08324.

[9] Thomas Steinke. Composition of differential privacy & privacy amplification by subsampling. `https://arxiv.org/abs/2210.00597`, oct 2022. Book chapter draft for *Differential Privacy for Artificial Intelligence Applications*.

[10] Yu-Xiang Wang. Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1905.00265*, 2019.