Differentially Private Linear Regression with Linked Data

Shurong Lin[†,*], Elliot Paquette[‡], Eric D. Kolaczyk[‡]
[†] Department of Mathematics and Statistics, Boston University
[‡] Department of Mathematics and Statistics, McGill University

Abstract. There has been increasing demand for establishing privacy-preserving methodologies for modern statistics and machine learning. Differential privacy, a mathematical notion from computer science, is a rising tool offering robust privacy guarantees. Recent work focuses primarily on developing differentially private versions of individual statistical and machine learning tasks, with nontrivial upstream pre-processing typically not incorporated. An important example is when record linkage is done prior to downstream modeling. Record linkage refers to the statistical task of linking two or more datasets of the same group of entities without a unique identifier. This probabilistic procedure brings additional uncertainty to the subsequent task. In this paper, we present two differentially private algorithms for linear regression with linked data. In particular, we propose a noisy gradient method and a sufficient statistics perturbation approach for the estimation of regression coefficients. We investigate the privacy-accuracy tradeoff by providing finite-sample error bounds for the estimators, which allows us to understand the relative contributions of linkage error, estimation error, and the cost of privacy. The variances of the estimators are also discussed. We demonstrate the performance of the proposed algorithms through simulations and an application to synthetic data.

Keywords: differential privacy, record linkage, data integration, privacy-preserving record linkage, gradient descent

Media Summary

Differential privacy is a mathematical framework for ensuring the privacy of individuals in datasets. It mitigates the privacy risk of disclosing sensitive information about individuals within the dataset during data analysis. Under such a framework, we are interested in finding the relationship between two variables (via statistical regression) *after* they are linked from two data sources with uncertainties. A pre-processing procedure of linking datasets is called record linkage, and the uncertainties should be taken into account in the downstream analysis. In the article, we propose two algorithms that satisfy differential privacy for regression estimation problems with linked data. The theoretical results regarding privacy guarantees and statistical accuracy are provided. We demonstrate the performance of the proposed algorithms through simulations and an application.

[*] shrlin@bu.edu. The research was supported in part by the U.S. Census Bureau Cooperative Agreement CB20ADR0160001 and Canadian NSERC RGPIN-2023-03566. The authors would like to thank Adam Smith (Boston University) for all the helpful discussions and comments.

1. Introduction

Data for the same group of entities are often scattered across different resources, lacking unique identifiers for perfect linkage. To conduct statistical modeling or inference on the integrated information, it is necessary to probabilistically link multiple datasets by comparing the common quasi-identifiers (e.g., names, gender, address) as a pre-processing step. Such a procedure is called record linkage (RL), also known as entity resolution, or data matching (Christen, 2012), which is an essential component of data integration in big data analytics (Dong & Srivastava, 2015). Thanks to its wide application in many disciplines such as public health and official statistics, record linkage has been studied for decades. Earlier pioneering works include Fellegi and Sunter (1969), Jaro (1989), and Newcombe et al. (1959). In addition, record linkage is frequently used in current practice. The U.S. Census Bureau has a long tradition using record linkage methodology for multiple endeavors. A current prominent example is the Decennial Census (U.S. Census Bureau, 2022). In this context, record linkage involves using administrative records and other data sources to improve data quality, with efforts underway to construct a comprehensive "reference database" including individuals from multiple administrative records. A recent review paper, Binette and Steorts (2022), provided a comprehensive summary of record linkage. Broadly speaking, there are two perspectives regarding record linkage (Chambers et al., 2021): (1) the primary viewpoint concerns how to link the records; (2) the secondary perspective is focused on how to propagate the uncertainty to the downstream statistical learning tasks *after* the linkage has been determined. Our focus in this paper will adopt the second of these two perspectives.

Closely related to record linkage is data privacy. In the area of privacy-preserving record linkage (PPRL), two or more private datasets owned by different organizations are linked without revealing the data to one another (Christen et al., 2020; Hall & Fienberg, 2010). The outcome of PPRL is the information regarding which pairs or sets are matched. PPRL, in turn, is associated with secure multiparty computation (SMPC) in that SMPC techniques are commonly used to solve PPRL problems (He et al., 2017; Kuzu et al., 2013; Rao et al., 2019). PPRL to date only engages in the private linkage process from a primary perspective, without concerning how the linkage uncertainties would impact the downstream analysis. On the contrary, the secondary perspective is to modify the statistical tools to account for the linkage uncertainty. Our goal is to incorporate privacy into the secondary perspective of record linkage, which is different from yet complementary to PPRL or SMPC.

Privacy concerns have, if anything, become significantly more exacerbated with the emergence of individual-level big data. Releasing information about a sensitive dataset is subject to a variety of privacy attacks (Dwork et al., 2017). Therefore, there has been a growing demand for establishing robust privacy-preserving methodologies for modern statistics and machine learning. A mathematical framework proposed by Dwork et al. (2006), differential privacy (DP), is now considered the gold standard for rigorous privacy protection and has made its way to broad application in industry (Apple, 2017; Google, 2021; Microsoft, 2020) and the public sector (U.S. Census Bureau, 2021). The literature on differential privacy has been flourishing in recent years and the interface of differential privacy and statistics has started to draw increasing attention from the statistics community.

Recent work on differential privacy focuses primarily on individual statistical and machine learning tasks, with nontrivial upstream pre-processing, such as record linkage, typically not incorporated. In this paper, we consider the linear regression problem, i.e.,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \mathcal{N}(0, \sigma^2 I_n) \tag{1.1}$$

but where X and $\boldsymbol{y}$ are observed in two separate datasets. As a result, rather than having X and $\boldsymbol{y}$ in hand, we are instead provided with a pair X and $\boldsymbol{z}$. Here $\boldsymbol{z}$ is a permutation of $\boldsymbol{y}$ resulting from record linkage performed by an external entity, who also supplies a minimum amount of information about the linkage accuracy. In the regression procedure, we take into account the linkage uncertainty as well as offer differential privacy guarantees. As shown in Figure 1 which depicts the pipeline of the problem we consider, we assume that an external analyst conducts record linkage a priori. From there, we aim to devise a private estimator for the regression coefficients of ultimate interest with the help of differential privacy.

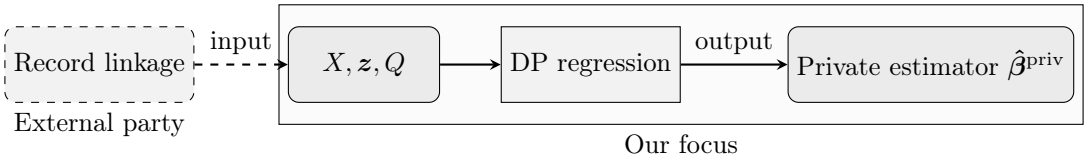


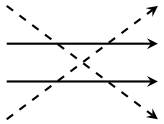
Figure 1. Pipeline of private regression with linked data.

Specifically, we propose two algorithms for linear regression after record linkage to meet differential privacy: (1) post-RL noisy gradient descent (NGD), and (2) post-RL sufficient statistics perturbation (SSP). Our work builds on the seminal work by Lahiri and Larsen (2005) where an estimator is proposed for linear regression with linked data in a non-privacy-aware setting. We construct a private estimator, $\hat{\boldsymbol{\beta}}^{\text{priv}}$, by deploying differential privacy tools to achieve privacy protections. To the best of our knowledge, our work is the first one in the literature to consider a statistical model after record linkage in a privacy-aware setting.

The two proposed algorithms also extend the noisy gradient method (Bassily et al., 2014) and the "Analyze Gauss" algorithm (Dwork et al., 2014), which are applied to linear regression, to additionally handle the presence of linkage errors. Prior works (Alabi et al., 2022; Bernstein & Sheldon, 2019; Cai et al., 2021; Sheffet, 2017; Wang, 2018) on differentially private linear regression do not consider possible record linkage pre-processing. If the data are linked beforehand, directly applying their algorithms to the imperfectly linked data is not ideal. It is well known that overlooking the linkage errors leads to substantial bias even with a high linkage accuracy (Neter et al., 1965; Scheuren & Winkler, 1993). Figure 2 showcases a toy example of record linkage, where mismatches, if treated as true, change the sign of the slope estimate. Our illustrative application later in the paper confirms this, where around 90% of the records are correctly linked, and the estimators ignoring linkage errors end up with large biases.

first name	last name	x
Shurong	Lin	1
Erik	K	2
Elliot	P	3
S	Li	4

Dataset A



first name	last name	y
S	L	2
Eric	K	4
Eliot	P	6
Sharon	Li	8

Dataset B

Figure 2. A toy example of record linkage with mismatches (dashed links). The true dataset $(X, \boldsymbol{y})$ is $\{(1,2),(2,4),(3,6),(4,8)\}$, yielding a slop estimate $\hat{\beta}_1 = 2$, while the linked set $(X, \boldsymbol{z})$ is given by $\{(1,8),(2,4),(3,6),(4,2)\}$, yielding $\hat{\beta}_1 = -1.6$.

Accompanying the estimators resulting from our algorithms, we provide mean-squared error bounds under typical regularity assumptions and record linkage schemes. When no linkage errors are present (i.e., a special case in our scenario), our result in Theorem 4.4 improves upon the noisy gradient method proposed in Cai et al. (2021) by using zero-concentrated differential privacy (zCDP, Bun and Steinke (2016)) to enable tighter bounds on privacy cost (see Lemma 2.3). Additionally, we have presented (approximate) theoretical variances for $\hat{\boldsymbol{\beta}}^{\text{priv}}$ resulting from both proposed algorithms. There appear to be very few other works that have addressed the issue of uncertainty. Two that we are aware of are Alabi (2022), who provided confidence bounds for the univariate case, and Sheffet (2017), who provided confidence intervals dependent on differential privacy noise. Our work focuses on the multivariate case and appears to be the first to directly work on exact variances rather than relying on bounds.

The remainder of this paper is organized as follows. Section 2 provides preliminaries on linear regression with linked data and differential privacy. We propose our two algorithms in Section 3 and present the relevant theoretical results in Section 4. In Section 5, we conduct a series of simulation studies and an application to synthetic data. Section 6 concludes and discusses future work. Complete proofs of all theorems can be found in the supplementary materials.

2. Preliminaries

In this section, we review the background results of linear regression after record linkage upon which we build our work, and fundamental concepts from differential privacy. Related work on linear regression with linked data and record linkage with privacy awareness are discussed.

2.1. Linear Regression with Record Linkage.

Let (X, Φ_X) and $(\boldsymbol{y}, \Phi_{\boldsymbol{y}})$ be two datasets that refer to the same group of n entities, with unknown one-to-one correspondence. The quasi-identifiers Φ_X and $\Phi_{\boldsymbol{y}}$ are used to perform the linkage procedure. Let $(X, \boldsymbol{z})$ be the linked data where $\boldsymbol{z}$ is a permutation of $\boldsymbol{y}$. Consider the following model for $\boldsymbol{z}$:

$$\mathbb{P}(z_i = y_j) = q_{ij}, \quad i, j = 1, \ldots, n, \tag{2.1}$$

then $\sum_{j=1}^{n} q_{ij} = 1$ for all i and $\sum_{i=1}^{n} q_{ij} = 1$ for all j. Thus, q_{ii} is the probability of the ith record being linked correctly. Let $Q = (q_{ij})$, which we call the matching probability matrix (MPM), a doubly stochastic matrix. The matrix Q can be estimated, for example, through bootstrapping (Chipperfield, 2020; Chipperfield & Chambers, 2015). In some cases, estimating Q can require inference on only a single parameter (e.g., in the exchangeable linkage error (ELE) model described in Section 2.1.1).

For the fixed-design homoskedastic linear model (1.1), when inference is done after record linkage based on $(X, \boldsymbol{z})$, Lahiri and Larsen (2005) proposed an unbiased estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{RL}} = (W^\top W)^{-1} W^\top \boldsymbol{z}, \tag{2.2}$$

where $W = QX$. Let $\boldsymbol{w}_i$ be the i-th row vector of W, then $\boldsymbol{w}_i = \sum_{j=1}^n q_{ij}\boldsymbol{x}_j$. Note that $\mathbb{E}(z_i) = \boldsymbol{w}_i^\top \boldsymbol{\beta}$, where the expectation is taken over both linkage uncertainties and $\boldsymbol{y}$. Transforming X into W offers bias correction for regression estimation after record linkage.

In addition, the variance of $\hat{\boldsymbol{\beta}}^{\mathrm{RL}}$ is given by

$$\Sigma^{\mathrm{RL}} \overset{def}{=} \mathrm{Var}(\hat{\boldsymbol{\beta}}^{\mathrm{RL}}) = (W^\top W)^{-1} W^\top \Sigma_{\boldsymbol{z}} W (W^\top W)^{-1}, \tag{2.3}$$

where $\Sigma_{\boldsymbol{z}} \overset{def}{=} \mathrm{Var}(\boldsymbol{z})$. Lahiri and Larsen (2005) provide the following characterization of the first two moments of $\boldsymbol{z}$.

Lemma 2.1 (Theorem A.1, Lahiri and Larsen, 2005)**.** *Under the model described by (1.1) and (2.1), we have for $i, j = 1, \ldots, n$*

- $\mathbb{E}(z_i) = \boldsymbol{w}_i^\top \boldsymbol{\beta}$;
- $\mathrm{Var}(z_i) = \sigma^2 + \boldsymbol{\beta}^\top A_i \boldsymbol{\beta}$ *with* $A_i = \sum_{j=1}^n q_{ij}(\boldsymbol{x}_j - \boldsymbol{w}_i)(\boldsymbol{x}_j - \boldsymbol{w}_i)^\top$;
- $\mathrm{Cov}(z_i, z_j) = \boldsymbol{\beta}^\top A_{ij} \boldsymbol{\beta}$ *with* $A_{ij} = \sum_{u=1}^n \sum_{v \neq u}^n q_{iu} q_{jv} (\boldsymbol{x}_i - \boldsymbol{w}_u)(\boldsymbol{x}_j - \boldsymbol{w}_v)^\top$.

Note that $\Sigma_{\boldsymbol{z}}$ involves the true coefficients $\boldsymbol{\beta}$ and $\Sigma_{\boldsymbol{z}} = \sigma^2 I_d + h(\boldsymbol{\beta}, Q, X)$ where $h(\boldsymbol{\beta}, Q, X)$ is a function of $\boldsymbol{\beta}, Q, X$ as elaborated in Lemma 2.1. Compared to the covariance of $\boldsymbol{y}$, $\Sigma_{\boldsymbol{z}}$ has an additional component $h(\boldsymbol{\beta}, Q, X)$ due to the uncertainty of record linkage.

2.1.1. *Structural Schemes of MPM.* The matching probability matrix (MPM) Q is generally assumed to have a simple structure. Two schemes used commonly in the literature are as follows.

Blocking Scheme. It is assumed that the MPM is a block diagonal matrix, which means the true matches only happen within blocks. Blocking significantly reduces the number of pairs for comparison and allows scalable record linkage. This scheme is used in almost all real-world applications, and different methods for blocking have been developed (Christen, 2012; Christophides et al., 2020; Steorts et al., 2014).

Exchangeable Linkage Errors (ELE) Model. The ELE model (Chambers, 2009) assumes homogeneous linkage accuracy and errors:

$$\begin{aligned} &\mathbb{P}(\text{correct linkage}) = q_{ii} = \gamma, \\ &\mathbb{P}(\text{incorrect linkage}) = q_{ij} = \frac{1-\gamma}{n-1} \text{ for } i \neq j. \end{aligned} \tag{2.4}$$

The ELE model has been adopted in recent works, such as Chambers et al. (2021) and Chambers et al. (2023), for various estimation problems. Even though (2.4) may oversimplify the reality, it is a representative model for a secondary analyst who has minimum information about the linkage quality. When blocking is used, the homogeneous linkage accuracy assumption is imposed within individual blocks. In other words, it still allows heterogeneous linkage accuracy between blocks.

2.2. **Differential Privacy.** Let $\mathcal{X}$ be some data space, and $D, D' \in \mathcal{X}^n$ be two neighboring datasets of size n which only differ in one record. Such a relation is denoted by $D \sim D'$.

Definition 1 ((ϵ, δ)-DP, Dwork and Roth, 2014)**.** *For $\epsilon > 0, \delta \geq 0$, a randomized algorithm $A: \mathcal{X}^n \to \mathcal{R}$ is (ϵ, δ)-differentially private if, for all $D \sim D' \in \mathcal{X}^n$ and any $\mathcal{O} \subseteq \mathcal{R}$,*

$$\mathbb{P}(A(D) \in \mathcal{O}) \leq e^\epsilon \cdot \mathbb{P}(A(D') \in \mathcal{O}) + \delta. \tag{2.5}$$

The expression (2.5) controls the distance between the output distributions on two neighboring datasets through the privacy budget ϵ and δ. Intuitively, differential privacy ensures that D is not distinguishable from D' based on the outputs. Thus, ϵ should be small enough for the privacy level to be meaningful. Typically, $\epsilon \in (10^{-3}, 10)$ and $\delta = o(1/n)$.

Differential privacy enjoys the following properties that facilitate the construction of differentially private algorithms.

Proposition 2.1 (Basic composition, Dwork and Roth, 2014)**.** *If f_1 is (ϵ_1, δ_1)-DP and f_2 is (ϵ_2, δ_2)-DP, then $f := (f_1, f_2)$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.*

Proposition 2.2 (Post-processing, Dwork and Roth, 2014)**.** *If f is (ϵ, δ)-DP, for any deterministic mapping g that takes $f(D)$ as an input, then $g(f(D))$ is (ϵ, δ)-DP.*

Generally, a differentially private algorithm is constructed by adding random noise from a certain structured distribution, such as the Laplace or Gaussian distributions. A notion central to the amount of noise we add is the sensitivity of the estimation function we desire to release privately.

Definition 2 (ℓ_2-sensitivity)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be an algorithm. The ℓ_2-sensitivity of f is defined as*

$$\Delta_f = \max_{D \sim D' \in \mathcal{X}^n} \|f(D) - f(D')\|_2. \tag{2.6}$$

The sensitivity of a function characterizes how much the output would change if one record in the dataset changes. To achieve (ϵ, δ)-DP, the amount of noise we need depends on both the budget and the sensitivity. The Gaussian Mechanism is a canonical example that will be employed herein, which does just that.

Lemma 2.2 (Gaussian mechanism, Dwork and Roth (2014))**.** *Let $0 < \epsilon < 1$ and $\delta > 0$. For an algorithm f on the dataset D, the Gaussian Mechanism $A(\cdot)$ defined as*

$$A(D) := f(D) + u, \tag{2.7}$$

where $u \sim \mathcal{N}(0, 2\ln(1.25/\delta)(\Delta_f/\epsilon)^2)$, is (ϵ, δ)-DP.

Combining the basic composition rule and the Gaussian mechanism, for a sequence of functions $(f_1, f_2, \ldots, f_T)$, let

$$u_t \sim \mathcal{N}\left(0, \frac{2T^2\Delta_t^2 \ln(1.25T/\delta)}{\epsilon^2}\right),$$

where Δ_t is the ℓ_2-sensitivity of f_t. Then, $A := (f_1 + u_1, f_2 + u_2, \ldots, f_t + u_T)$ satisfies (ϵ, δ)-DP. However, as T increases, this construction tends to add more noise than necessary due to the loose composition. Instead, we could utilize zero-concentrated differential privacy (zCDP, Bun and Steinke (2016)), another variant of DP, to achieve tighter composition for (ϵ, δ)-DP. The following Lemma essentially captures the results from Bun and Steinke (2016), formulated for our purposes.

Lemma 2.3 (Better composition for (ϵ, δ)-DP via zCDP)**.** *Let $\epsilon > 0, \delta > 0$. For a sequence of functions $(f_1, f_2, \ldots, f_T)$, let*

$$u_t \sim \mathcal{N}\left(0, \frac{T\Delta_t^2}{2\rho}\right), \tag{2.8}$$

with $\rho := \epsilon + 2\ln(1/\delta) - 2\sqrt{(\epsilon + \ln(1/\delta))\ln(1/\delta)}$. *Then, the randomized algorithm* $A := (f_1 + u_1, f_2 + u_2, \ldots, f_T + u_T)$ *satisfies* (ϵ, δ)*-DP. If* $\epsilon \leq \frac{8\ln(1/\delta)}{2+\sqrt{2}}$, *it suffices to have*

$$u_t \sim \mathcal{N}\left(0, \frac{4T\Delta_t^2 \ln(1/\delta)}{\epsilon^2}\right). \tag{2.9}$$

Please refer to the supplementary materials for details. Since, in most practical budget settings, we have $\epsilon \leq \frac{8\ln(1/\delta)}{2+\sqrt{2}}$, we will apply (2.9) for composition and analysis in the rest of the paper, acknowledging that (2.8) is valid for all parameter ranges.

In Section 3, we shall employ Lemmas 2.2 and 2.3 in devising two distinct algorithms for linear regression after record linkage.

2.3. **Related Work.** Linear regression with linked data is a fundamental statistical task that has been explored in various articles. Scheuren and Winkler (1993) initially considered the linkage model (2.1) for linear regression and proposed an estimator that is not generally unbiased. Later, Lahiri and Larsen (2005) introduced an exactly unbiased OLS-like estimator given in (2.2) with an expression for the variance, which outperformed the approach by Scheuren and Winkler (1993). Besides, Chambers (2009) and Zhang and Tuoto (2021) offered a few other estimators. According to their simulation studies, some of the estimators provided performance that was at most similar, but not noticeably better, compared to the one proposed by Lahiri and Larsen (2005). Yet, Zhang and Tuoto (2021) relaxed the condition by not assuming that the probability of correct linkage, q_{ii} in the model (2.1), can be obtained or estimated. For more extensive reviews of this literature, Wang et al. (2022) gave an account of the recent development of various methods on regression analysis with linked datasets. Chambers et al. (2023) reviewed current research on robust regression of linked data.

On the other hand, there is ongoing research on privacy-preserving record linkage (PPRL) in the field of computer science. PPRL aims to privately link multiple sensitive datasets held by different organizations when they are unwilling or not permitted to share their data with external parties due to privacy and confidentiality concerns. To achieve privacy protection, techniques such as SMPC and DP are combined with machine learning and deep learning methods for conducting PPRL (Christen et al., 2020; Gkoulalas-Divanis et al., 2021; Ranbaduge et al., 2022). PPRL primarily concerns data leakage during the linkage process and produces a linked dataset that can be used for further analysis, yet most applications treat the linked data as if there were no linkage errors. Neither the uncertainty propagation nor private release of the downstream analysis is considered within the scope of PPRL.

Note that there are several articles on privacy-preserving analysis on vertically partitioned databases. In these databases, the attributes are distributed among multiple parties, but common unique identifiers exist to facilitate data linkage across the different parties. Unlike probabilistic record linkage, vertically partitioned databases do not involve linkage errors. Du et al. (2004), Gascón et al. (2017), Hall et al. (2011), and Sanil et al. (2004) discussed the implementations of privacy-preserving linear regression protocols that prevent data disclosure across organizations, whereas Dwork and Nissim (2004) considered data mining from the perspective of the private release of statistical querying in a spirit similar to our work.

3. Differentially Private Algorithms

The unbiased and simply structured estimator provided in (2.2) with a known closed-form variance makes it a suitable prototype to construct our private estimators. We introduce two differentially private algorithms in the following, based on (1) noisy gradient descent, and (2) sufficient statistics perturbation. As the names suggest, we mitigate privacy risk by perturbing either the gradient or sufficient statistics during the computation of the linear model. Hereafter, if not specified otherwise, $\|\cdot\|$ denotes the 2-norm.

3.1. Post-RL Noisy Gradient Descent.

Gradient descent methods are ubiquitous in scientific computing for numerous optimization problems. Within the framework of differential privacy, Bassily et al. (2014) provided a noisy variant of the classic gradient descent algorithm. It was later adapted by Cai et al. (2021) to solve the classic linear regression problem with faster convergence. Leveraging the work by Bassily et al. (2014) and Cai et al. (2021), we tailor the noisy gradient method for the post-RL linear regression model for $(X, \boldsymbol{z})$ based on (1.1) and (2.1).

Let $\mathcal{L}_n(\boldsymbol{\beta}) \overset{def}{=} \frac{1}{2n}(\boldsymbol{z}-W\boldsymbol{\beta})^\top(\boldsymbol{z}-W\boldsymbol{\beta})$ be the loss function, where recall $W = QX$. The minimizer of $\mathcal{L}_n(\boldsymbol{\beta})$ is the non-private RL estimator proposed by Lahiri and Larsen (2005). Let $\Pi_R(\boldsymbol{r})$ denote the projection of $\boldsymbol{r} \in \mathbb{R}^s$ onto the ℓ_2 ball $\{\boldsymbol{r} \in \mathbb{R}^s : \|\boldsymbol{r}\| \le R\}$. The post-RL noisy gradient descent (NGD) algorithm is defined as follows.

Algorithm 1 Post-RL Noisy Gradient Descent

Input: Linked dataset $(X, \boldsymbol{z})$ and matching probability matrix Q, privacy budget (ϵ, δ), noise scale factor B, step size η, number of iterations T, truncation level R, feasibility C, initial value $\boldsymbol{\beta}^0$.

1: Let $W = QX$.

2: **for** $t = 0$ to $T-1$ **do**

3: Generate $\boldsymbol{u}_t \sim \mathcal{N}\left(0, \omega^2 I_d\right)$ where $\omega = \dfrac{2\eta B\sqrt{T\ln(1/\delta)}}{n\epsilon}$.

4: Compute

$$\boldsymbol{\beta}^{t+1} = \Pi_C(\boldsymbol{\beta}^t - \frac{\eta}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^\top\boldsymbol{\beta}^t - \Pi_R(z_i))\boldsymbol{w}_i + \boldsymbol{u}_t). \tag{3.1}$$

5: **end for**

Output: $\hat{\boldsymbol{\beta}}^{priv} = \boldsymbol{\beta}^T$.

Algorithm 1 is a modified version of the projected gradient descent that incorporates (1) post-RL transformation of the design matrix, (2) addition of noise $\boldsymbol{u}_t$ at each gradient step, and (3) use of projection $\Pi_R(\cdot)$ on the response variable. The regular parameters, including η, T and C for the projected gradient method, are specified in Theorem 4.4 for the discussion of the accuracy of $\hat{\boldsymbol{\beta}}^{\text{priv}}$. The injection of noise follows Lemma 2.3. The scale of the Gaussian noise $\boldsymbol{u}_t$ at step t depends on the privacy budget (ϵ, δ), and the noise scale factor B associated with the sensitivity in the update function (3.1). The purpose of the projection on $\boldsymbol{z}$ is to bound the sensitivity of the gradient. With a proper choice of R that scales up with $\sqrt{\ln n}$ (specified in Section 4), the projection does not affect the accuracy of the final estimator with high probability.

The major challenge lies in calculating the sensitivity. In the non-RL least square regression, two neighboring datasets $D = (X, \boldsymbol{y})$ and $D = (X', \boldsymbol{y}')$ differ in a single row, making it straightforward to derive the sensitivity of the gradient of $\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{y} - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta})$. Here, in the context

of post-RL analysis, we consider two neighboring datasets containing both linking variables and regression variables, denoted as $D = (X, \Phi_X, \boldsymbol{y}, \Phi_{\boldsymbol{y}})$ and $D' = (X', \Phi_{X'}, \boldsymbol{y}', \Phi_{\boldsymbol{y}'})$, which differ in the record of one individual. The change in one row of the quasi-identifiers Φ_X and $\Phi_{\boldsymbol{y}}$ may affect more than one row of the matching probability matrix Q. As a result, the entries of the transformed design matrix $W = QX$ subject to change are not limited to one row as in the non-RL case. Consequently, determining the sensitivity of the gradient of $\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{z} - W\boldsymbol{\beta})^\top(\boldsymbol{z} - W\boldsymbol{\beta})$ becomes non-trivial. This challenge distinguishes our work from Cai et al. (2021). However, we will demonstrate in Section 4 that, under a condition on the structure of Q, the sensitivity can be tracked.

3.2. **Post-RL Sufficient Statistics Perturbation.** Noise can be injected into the process besides the gradient computation. Since the estimator interacts with the data through its (joint) sufficient statistics, an efficient way is to perturb the sufficient statistics to protect the data. Such a technique, sufficient statistics perturbation (SSP), has been used in previous works such as Foulds et al. (2016), Vu and Slavkovic (2009), and Wang (2018). For the non-private OLS estimator $\hat{\boldsymbol{\beta}}^{\text{OLS}} = (X^\top X)^{-1} X\boldsymbol{y}$, to perturb the joint sufficient statistics $(X^\top X, X\boldsymbol{y})$, it suffices to add noise to $A^\top A$ where $A = (X \mid \boldsymbol{y})$ is the augmented matrix. Dwork et al. (2014) offered an algorithm, "Analyze Gauss", to privately release $A^\top A$. It was later utilized by Sheffet (2017) for private linear regression, primarily perturbing the sufficient statistics.

In our work, we adapt the "Analyze Gauss" algorithm to linear regression after record linkage, as shown in Algorithm 2. The noise scale factor B is the sensitivity of $A^\top A \overset{def}{=} \begin{pmatrix} W^\top W & W^\top \boldsymbol{z} \\ \boldsymbol{z}^\top W & \boldsymbol{z}^\top \boldsymbol{z} \end{pmatrix}$ which is specified in Section 4. The gram matrix $A^\top A$ exhibits properties that facilitate the computation of its sensitivity. Algorithm 2 illustrates how incorporating the joint sufficient statistics in a comprehensive form facilitates the deployment of differential privacy.

Algorithm 2 Post-RL Sufficient Statistics Perturbation

Input: Linked dataset $(X, \boldsymbol{z})$ and matching probability matrix Q, privacy budget (ϵ, δ), noise scale factor B, truncation level R.

1: Let $W = QX$.
2: Generate a $d \times d$ symmetric Gaussian random matrix U whose upper triangle entries (including the diagonal) are sampled i.i.d. from $\mathcal{N}(0, \omega^2)$ where $\omega = \dfrac{B\sqrt{2\ln(1.25/\delta)}}{\epsilon}$.
3: Generate a d-dimensional Gaussian random vector $\boldsymbol{u}$ whose entries are sampled i.i.d. from $\mathcal{N}(0, \omega^2)$.
4: **if** $W^\top W + U$ is computationally singular **then**
5: Repeat steps $2 \sim 3$.
6: **end if**

Output: $\hat{\boldsymbol{\beta}}^{\text{priv}} = (W^\top W + U)^{-1}(W^\top \boldsymbol{z}^* + \boldsymbol{u})$ where $z^* = (\Pi_R(z_1), \ldots, \Pi_R(z_n))^\top$.

Remark 3.1. *In step 4, by post-processing, checking for singularity of $W^\top W + U$ consumes no extra privacy budget. In fact, the probability of $W^\top W + U$ being singular decreases exponentially as the sample size increases.*

An alternative approach to implementing the SSP method is to add random noise separately to each sufficient statistic. In this approach, the total privacy budget should be divided between $X^\top X$

and $X\boldsymbol{y}$ for the estimation of linear regression, as proposed by Wang (2018). However, treating the joint statistics as a whole is more economical in terms of budgeting in general. Lin et al. (2023) showed through comparison that splitting the total budget among the components results in introducing larger noise on average. Although adding noise individually to the components of interest allows for the private release of each quantity, it is not part of the goal of the estimation.

4. Theoretical Results

In this section, we provide the theoretical results of the two algorithms introduced in Section 3. The results are threefold: (1) differential privacy guarantees, (2) finite-sample error bounds, and (3) variances of the private estimators. We present each of these along with a discussion of the corresponding conditions as they relate to the main variables in our record linkage model. All proofs for these results can be found in the supplementary materials.

4.1. Privacy Guarantees.

The algorithms are designed to achieve certain privacy guarantees, given the corresponding sensitivity, for the post-RL case:

Theorem 4.1 (Privacy Guarantees)**.** *Assume the following boundedness conditions hold:*

(A1) There is a constant $c_x < \infty$ such that $\|\boldsymbol{x}\|_2 \leq c_x$.

(A2) Let Q and Q' be the matching probability matrices (MPMs) resulting from the neighboring datasets D and D' and let $Q \sim Q'$ denote such a relation. We assume that $\sup_{Q\sim Q'} \|Q-Q'\|_1 \leq M$ for some constant $M < \infty$, where $\|\cdot\|_1$ is the entry-wise 1-norm.

Given the linked data $(X, \boldsymbol{z})$ and the matching probability matrix Q for the regression problem in (1.1), under Assumptions (A1) and (A2), it follows that

(1) Algorithm 1 satisfies (ϵ, δ)-differential privacy with

$$B = Rc_x(M+4) + 2Cc_x^2(M+2), \tag{4.1}$$

(2) Algorithm 2 satisfies (ϵ, δ)-differential privacy with

$$B = Rc_x(M+4) + \max\{2c_x^2(M+2), 2R^2\}. \tag{4.2}$$

Essentially, we assume that the data domain is bounded, which is critical for deriving a finite sensitivity of the target function on the data. (A1) is a standard assumption for a *bounded design* X. For the linking variables that are generally categorical, there are no analogous definitions of "norm" for numerical vectors. Instead, (A2) is imposed on the MPM since it summarizes all the information of the linking variables in the linkage model we consider. Specifically, we assume that two MPMs produced by two neighboring datasets do not differ much in terms of the entry-wise 1 norm. This assumption characterizes a *bounded linkage model*.

The rationale of (A2) is supported by typical schemes imposed on the structures of MPM in practice, as reviewed in Section 2.1.1. For example, with the blocking scheme, the size of each block is manageably small (O(1)). When one record is altered, the fluctuation of the MPM is limited to at most two blocks. Additionally, with the ELE model (2.4), as long as the changes to a single record only affect a finite number of records, the linkage accuracy γ changes at most $O(1/n)$. Therefore, we have $\sup_{Q\sim Q'} \|Q - Q'\|_1 = O(1)$. In general, a robust record linkage approach should not produce two considerably different MPMs from two neighboring datasets. Therefore, it is realistic to assume a bounded linkage model.

The proofs of Theorem 4.1 revolve around calculating the sensitivity of the target function in each algorithm. Besides the upper bounds c_x and M discussed above, the sensitivity also depends on the truncation level R on the response. Truncation is commonly used in DP algorithm designs when there are no priori bounds on the relevant quantities (e.g., Abadi et al. (2016)). In Section 4.2, we provide a specific choice of R and present an accuracy statement with high probability.

4.2. **Finite-Sample Error Bounds.** We study the accuracy of the proposed estimators by deriving the finite-sample error bounds. In the following, we introduce two more assumptions in addition to (A1) and (A2):

(A3) The true parameter $\boldsymbol{\beta}$ satisfies $\|\boldsymbol{\beta}\|_2 \le c_0$ for some constant $0 < c_0 < \infty$.

(A4) The minimum and maximum eigenvalues of $W^\top W/n$ satisfy

$$0 < \frac{1}{L} < d\lambda_{\min}\left(\frac{W^\top W}{n}\right) \le d\lambda_{\max}\left(\frac{W^\top W}{n}\right) < L \tag{4.3}$$

for some constant $1 < L < \infty$.

Assumption (A4) implies the smoothness and strong convexity of the loss function $\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{z} - W\boldsymbol{\beta})^\top(\boldsymbol{z} - W\boldsymbol{\beta})$, which allows for a fast convergence rate for the gradient descent method in Algorithm 1. On the other hand, for Algorithm 2, note that the term $(W^\top W)^{-1}$ is a component of sufficient statistics. Assumption (A4) offers a bound on the norm of $(W^\top W)^{-1}$, which helps derive the error bound of $\hat{\boldsymbol{\beta}}^{\text{priv}}$. Let Assumption (A4') be (A4) with W replaced by X and the constant L replaced by L'. The larger of L and L' can be chosen as the constant to satisfy both (A4) and (A4'). Therefore, for convenience, we consider (A4) and (A4') to be the same assumption. We first obtain the accuracy of the non-private estimators, for comparison purposes.

Lemma 4.2. *Let* $\hat{\boldsymbol{\beta}}^{OLS} = \arg\min_{\boldsymbol{\beta}}(\boldsymbol{y} - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta})$ *be the OLS estimator. Then, under (A4), it follows that* $\mathbb{E}\|\hat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\|^2 = \sigma^2 \operatorname{tr}(X^\top X)^{-1} = \Theta\left(\frac{\sigma^2 d^2}{n}\right)$.

Lemma 4.3. *Let* $\hat{\boldsymbol{\beta}}^{RL} = \arg\min_{\boldsymbol{\beta}}(\boldsymbol{z} - W\boldsymbol{\beta})^\top(\boldsymbol{z} - W\boldsymbol{\beta})$ *be the non-private record linkage estimator, and* Σ^{RL} *be the covariance matrix of* $\hat{\boldsymbol{\beta}}^{RL}$*. Then,*

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{RL} - \boldsymbol{\beta}\|^2 = \operatorname{tr}(\Sigma^{RL}), \tag{4.4}$$

where $\Sigma^{RL} = (W^\top W)^{-1} W^\top \Sigma_{\boldsymbol{z}} W (W^\top W)^{-1}$.

As a special case, when the linkage is perfect (i.e., Q is an identity matrix), the expected error of $\hat{\boldsymbol{\beta}}^{\text{RL}}$ in (4.4) takes the reduced form $\sigma^2 \operatorname{tr}(X^\top X)^{-1}$ which is exactly the lower bound obtained by $\hat{\boldsymbol{\beta}}^{\text{OLS}}$. Then, by Lemma 4.2, we know that $\mathbb{E}\|\hat{\boldsymbol{\beta}}^{\text{RL}} - \boldsymbol{\beta}\|^2$ is of order at least $\dfrac{\sigma^2 d^2}{n}$ under (A4). From a secondary perspective regarding record linkage, it is beyond our scope to study how $\operatorname{tr}(\Sigma^{\text{RL}})$ behaves in general.

For the two proposed algorithms, we present upper bounds of the excess squared error of the private estimators, namely, $\|\hat{\boldsymbol{\beta}}^{\text{priv}} - \hat{\boldsymbol{\beta}}^{\text{RL}}\|^2$.

Theorem 4.4 (Post-RL NGD)**.** *Given the linked data* $(X, \boldsymbol{z})$ *and the matching probability matrix* Q *for the regression problem in (1.1), set the parameters of Algorithm 1 as follows:*

- *step size* $\eta = d/L$*, number of iterations* $T = \lceil L^2 \ln(c_0^2 n) \rceil$*, feasibility* $C = c_0$*, initialization* $\boldsymbol{\beta}^0 = \mathbf{0}$*;*

- *truncation level* $R = \sigma\sqrt{2\ln n}$;
- *noise scale factor* $B = Rc_x(M+4) + 2c_0c_x^2(M+2)$;

Under Assumptions (A1)-(A4), given $\delta = o(1/n)$, *with probability at least* $1 - c_1e^{-c_2\ln n} - e^{-c_3 d}$ *where* c_1, c_2, c_3 *are constants (see the proof), it follows that*

$$\|\hat{\boldsymbol{\beta}}^{priv} - \hat{\boldsymbol{\beta}}^{RL}\|^2 = \frac{1}{n} + O\left(\frac{\sigma^2 d^3 \ln^2 n \ln(1/\delta)}{n^2\epsilon^2}\right). \tag{4.5}$$

Theorem 4.5 (Post-RL SSP)**.** *Given the linked data* $(X, \boldsymbol{z})$ *and the matching probability matrix* Q *for the regression problem in (1.1), in Algorithm 2, set*

- *truncation level* $R = \sigma\sqrt{2\ln n}$;
- *noise scale factor* $B = Rc_x(M+4) + 2\max\{c_x^2(M+2), R^2\}$.

Under Assumptions (A1)-(A4), given $\delta = o(1/n)$, *with probability at least* $1 - c_1e^{-c_2\ln n} - e^{-c_3 d}$ *where* c_1, c_2, c_3 *are constants (see the proof),*

$$\|\hat{\boldsymbol{\beta}}^{priv} - \hat{\boldsymbol{\beta}}^{RL}\|^2 = O\left(\frac{\sigma^4 d^3 \ln^2 n \ln(1/\delta)}{n^2\epsilon^2}\right). \tag{4.6}$$

In both algorithms, the response is projected with a level $R = \sigma\sqrt{2\ln n}$ where σ^2 is the homoskedastic variance of the random error in linear model (1.1). Let $\mathcal{E} = \{\Pi_R(z_i) = z_i, \forall i \in [n]\}$, then $\mathcal{E}$ is a high-probability event. The error bound is analyzed under $\mathcal{E}$, thus we obtain a statement with high probability.

In the NGD method, the bound consists of two parts on the RHS in (4.5). The first error term $1/n$ results from the convergence rate of gradient descent after T iterations. The second error term is due to the addition of Gaussian noise for privacy and thus involves ϵ, δ. It is worth noting that the choice in theory $T = \lceil L^2 \ln(c_0^2 n)\rceil$ is, to some extent, conservative to ensure the first error term is $O(1/n)$, which is the same order as $\mathbb{E}\|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}\|^2$. However, more iterations give rise to larger random noise being added to gradient updates due to a smaller privacy budget per iteration. In practice, a smaller number of iterations may be favored for the tradeoff (see the experiment in Section 5.2), especially when n is not sufficiently large.

For the SSP algorithm, the convergence rate in (4.6) depends on similar variables as in the NGD algorithm. The major difference is that it is controlled by σ^4 instead of σ^2 due to the sensitivity of the gram matrix $A^\top A$ defined in Section 3.2. However, the SSP method has a faster convergence rate when n is sufficiently large. As a result, the SSP estimator is more susceptible to a large variance of the random error in the response variable whereas the NGD method is more robust. As we shall see in Section 5, the performance of the two algorithms is different under various scenarios.

Putting together Lemma 4.3 and Theorems 4.4 and 4.5, we obtain a high probability error bound for each algorithm as follows.

Corollary 4.1. *Under the regularity conditions (A1)-(A4),*

(i) (Post-RL NGD)

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{priv} - \boldsymbol{\beta}\|^2 = O\left(\mathrm{tr}(\Sigma^{RL}) + \frac{\sigma^2 d^3 \ln^2 n \ln(1/\delta)}{n^2\epsilon^2}\right) \tag{4.7}$$

with probability at least $1 - c_1e^{-c_2\ln n} - e^{-c_3 d}$.

(ii) (Post-RL SSP)

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{priv} - \boldsymbol{\beta}\|^2 = O\left(\mathrm{tr}(\Sigma^{RL}) + \frac{\sigma^4 d^3 \ln^2 n \ln(1/\delta)}{n^2\epsilon^2}\right) \tag{4.8}$$

with probability at least $1 - c_1 e^{-c_2 \ln n} - e^{-c_3 d}$.

4.3. **Variances.** As discussed in the Introduction, although a few works (Alabi, 2022; Sheffet, 2017) have addressed uncertainty of DP estimators through confidence bounds and intervals, the exact variance of DP estimators is rarely determined in most cases. Recent work, such as Lin et al. (2023), has explored the variance of the private estimators for population proportions that have fairly simple structures. The main barrier to the inspection of variance is that if the noise is injected into the intermediate steps of the estimation process other than the output, then it is difficult to track the variability that noise introduces to the output estimator due to the intricate nature of the algorithm.

The NGD and SSP algorithms are two examples where noise is added in the middle of the estimation process. The operations like function composition and taking the inverse complicate the inspection of the variance of the output estimator $\hat{\boldsymbol{\beta}}^{\mathrm{priv}}$. To address this issue, we investigate the variance of $\hat{\boldsymbol{\beta}}^{\mathrm{priv}}$ for the two algorithms by studying the variances of two proxy estimators. The theoretical variances of the proxy estimators can be used to approximate those of $\hat{\boldsymbol{\beta}}^{\mathrm{priv}}$.

Theorem 4.6 (Variance for Post-RL NGD)**.** *In Algorithm 1 , if we consider the estimator without projections*

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \frac{\eta}{n} W^\top (W\boldsymbol{\beta}^t - \boldsymbol{z}) + \boldsymbol{u}_t, \tag{4.9}$$

then the variance of the T*th iterate is given by*

$$\Sigma = \sum_{t=1}^{T} (I_d - A)^{t-1} \cdot B^\top \Sigma_{\boldsymbol{z}} B \cdot \sum_{t=1}^{T} (I_d - A)^{t-1} + \omega^2 \sum_{t=1}^{T} (I_d - A)^{2t-2}, \tag{4.10}$$

where I_d *is the identity matrix of size* d*,* $A \overset{def}{=} \frac{\eta}{n} W^\top W$*,* $B \overset{def}{=} \frac{\eta}{n} W$*, and* ω^2 *is the variance of* $\boldsymbol{u}_t$*.*

Remark 4.1. *In the non-private case where* $\omega^2 = 0$*, let* $T \to \infty$*, in which case*

$$\Sigma \to A^{-1} B^\top \Sigma_{\boldsymbol{z}} B A^{-1} = (W^\top W)^{-1} W^\top \Sigma_{\boldsymbol{z}} W (W^\top W)^{-1} = \Sigma^{RL},$$

which is exactly the variance of $\hat{\boldsymbol{\beta}}^{RL}$ *given in (2.3).*

The estimator in Algorithm 1 is a projected variant of (4.9). The use of projection with level C on $\boldsymbol{\beta}^{t+1}$ in (3.1) impedes the exact analysis of variance for $\hat{\boldsymbol{\beta}}^{\mathrm{priv}}$. Instead, we provide the variance in (4.10) for the non-projected estimator as a conservative variance for $\hat{\boldsymbol{\beta}}^{\mathrm{priv}}$. The level of projection, the scale of noise, and the number of iterations together determine how conservative it is. From Remark 4.1, we know that as T increases, the first term in the RHS of (4.10) is getting close to Σ^{RL}. The second term, $\omega^2 \sum_{t=1}^{T} (I_d - A)^{2t-2}$, then summarizes the cumulative variability resulting from adding random noise at each iteration. Note that this term does not converge by simply increasing T, due to the fact that a smaller budget leads to larger noise at each iteration.

Theorem 4.7 (Variance for Post-RL SSP). *For Algorithm 2, let* $\hat{\boldsymbol{\beta}}' = \hat{\boldsymbol{\beta}}^{RL} + (W^\top W)^{-1}\boldsymbol{u} - (W^\top W)^{-1} \cdot U(\hat{\boldsymbol{\beta}}^{RL} + (W^\top W)^{-1}\boldsymbol{u})$, *then* $\hat{\boldsymbol{\beta}}^{priv} - \hat{\boldsymbol{\beta}}' \xrightarrow{p} 0$ *as* $n \to \infty$. *The variance of* $\hat{\boldsymbol{\beta}}'$ *is given by*

$$\Sigma = \Sigma^{RL} + \omega^2 (W^\top W)^{-1}(I_d + \Sigma_0 + \Sigma_1 + \Sigma_2)(W^\top W)^{-1}, \tag{4.11}$$

where $\Sigma^{RL} = \mathrm{Cov}(\hat{\boldsymbol{\beta}}^{RL})$ *and the entries of* Σ_0, Σ_1 *and* Σ_2 *are given by*

- $(\Sigma_0)_{kk} = \sum_{i=1}^d \beta_i^2$ *for* $k = 1, ..., d$; $(\Sigma_0)_{kl} = \beta_k \beta_l$ *for* $k \neq l$.
- $(\Sigma_1)_{kk} = \sum_{i=1}^d \Sigma_{ii}^{RL}$ *for* $k = 1, ..., d$; $(\Sigma_1)_{kl} = \Sigma_{kl}^{RL}$ *for* $k \neq l$.
- $(\Sigma_2)_{kk} = \sum_{i=1}^d \Sigma'_{ii}$ *for* $k = 1, ..., d$; $(\Sigma_2)_{kl} = \Sigma'_{kl}$ *for* $k \neq l$, *where* $\Sigma' \overset{def}{=} \omega^2 (W^\top W)^{-2}$.

Remark 4.2. *As shown in the proof of 4.7 (see the supplemental), the proxy estimator* $\hat{\boldsymbol{\beta}}'$ *is a first-order approximation for* $\hat{\boldsymbol{\beta}}^{priv}$ *using Taylor series for the term* $(I + U(W^\top W)^{-1})^{-1}$ *which appears in the decomposition of* $\hat{\boldsymbol{\beta}}^{priv}$.

The variance of $\hat{\beta}'$ also consists of two parts: the variance of the non-private estimator $\hat{\boldsymbol{\beta}}^{\mathrm{RL}}$ and the additional variation due to the noise injected for privacy purposes. Given Assumption (A4), we have $\|(W^\top W)^{-1}\| = O(d/n)$ that appears in Σ_1 and Σ_2. As n increases, the dominant component of the second term would be $\omega^2 (W^\top W)^{-1}(I_d + \Sigma_0)(W^\top W)^{-1}$.

5. Numerical Results

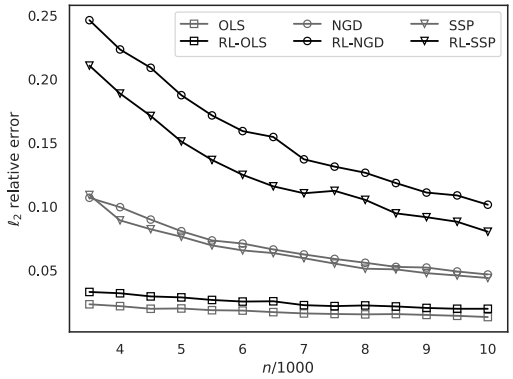
To evaluate the finite-sample performance of the proposed algorithms, we conduct a series of simulation studies and an application to a synthetic dataset that contains real data.

5.1. Simulation Studies.

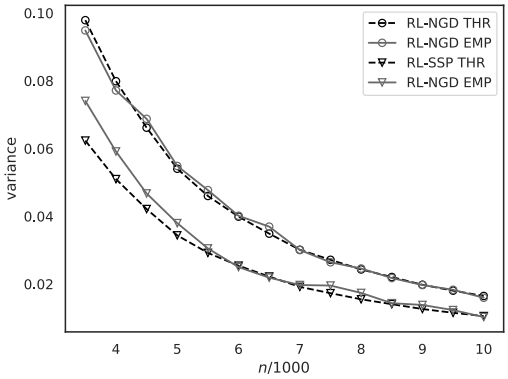
In this section, we conduct simulation studies to assess the performance of the two proposed algorithms for simple linear regression with linked data. The non-private OLS estimator and RL estimator $\hat{\boldsymbol{\beta}}^{\mathrm{RL}}$ by Lahiri and Larsen (2005) are included as benchmarks. The private, non-RL counterpart methods are also performed in the absence of linkage errors for comparison. For each simulation, a fixed design matrix X and an matching probability matrix Q are produced and a total of 1000 repetitions are run over the randomness of both the intrinsic error $\boldsymbol{e} \sim \mathcal{N}(0, \sigma^2 I_n)$ of the regression model and the noise injected for privacy. Figure 3 displays the ℓ_2 relative error and both empirical and theoretical variances for the two settings.

Two sets of simulations are conducted to explore the performance with varying sample size n and σ, the homoskedastic variance of the random error in linear model (1.1). The parameters are set as follows:

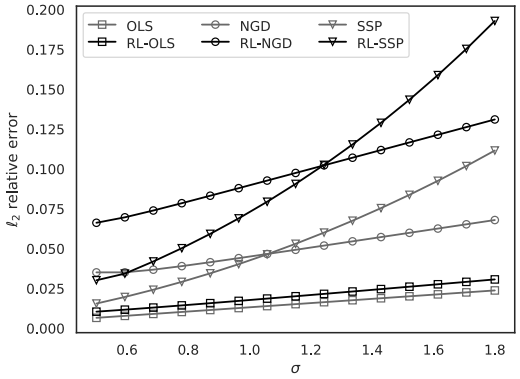
- ELE linkage model: blockwise linkage accuracy γ_i characterizing Q, block size $n_i = 25$.
  - Settings 1 and 2: $\gamma_i \in \mathrm{uniform}(0.6, 0.9)$, $M = 1$ in Assumption (A2).
  - Setting 3: the linkage accuracy $\gamma_i \equiv \gamma$ which varies from 0.6 to 1, while M scales from 1 to 0.
- regression model: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathrm{uniform}(-1, 1)$, true regression coefficient $\beta = 1$.
  - Setting 1: n varies from 3,000 to 10,000, σ is fixed at 1.
  - Setting 2: n is fixed at $10,000$, σ varies from 0.5 to 1.8.
  - Setting 3: n is fixed at $10,000$, σ is fixed at 1.
- privacy budget: $\epsilon = 1$, $\delta = 1/n^{1.1}$.



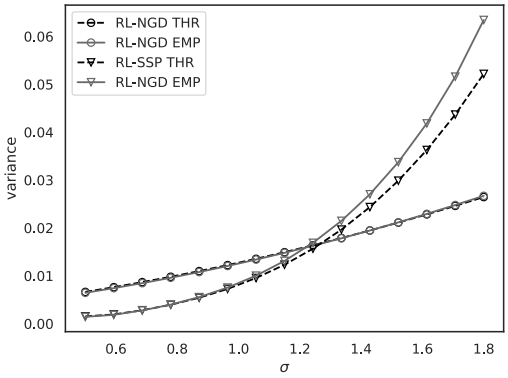
(A) Setting 1: $\sigma = 1$



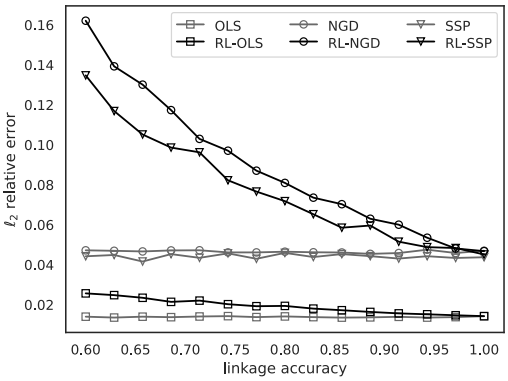
(B) Setting 1: $\sigma = 1$



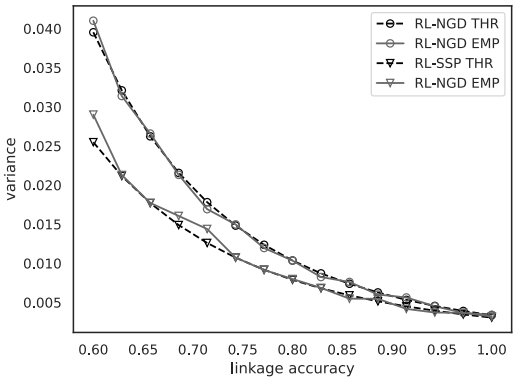
(C) Setting 2: $n = 10,000$



(D) Setting 2: $n = 10,000$



(E) Setting 3: $\sigma = 1$, $n = 10,000$



(F) Setting 3: $\sigma = 1$, $n = 10,000$

Figure 3. Average ℓ_2-error and variance (theoretical versus empirical), with $(\epsilon, \delta) = (1, 8.5\times 10^{-5})$, against n and σ, respectively.
The "RL-NGD" and "RL-SSP" algorithms are our proposed post-RL approaches applied to the linked data, compared with the non-RL "NGD" and "SSP" methods applied to $(X, \boldsymbol{y})$ (i.e., with no linkage errors). The non-private "OLS" and "RL-OLS" (Lahiri & Larsen, 2005) results are also plotted for benchmarking. The number of iterations for "RL-NGD" results fall within the range of $(210, 260)$.

In setting 1, where σ is fixed at 1, Figure 3a shows the errors of all methods decrease with a growing sample size. Due to the linkage errors, the post-RL methods, including $\hat{\boldsymbol{\beta}}^{\mathrm{RL}}$ and our two algorithms (denoted as "RL-OLS", "RL-NGD", and "RL-SSP" in the figures) run on the linked data $(X, \boldsymbol{z})$, naturally always yield larger errors than their counterparts run on $(X, \boldsymbol{y})$ when no linkage has to be done beforehand. In this case, with $\sigma = 1$, post-RL SSP outperforms post-RL NGD in terms of both accuracy and variance. However, as σ increases, post-RL NGD algorithm starts perform better, as depicted in Figure 3c with varying σ. The error grows linearly for post-RL NGD and quadratically for post-RL SSP, which aligns with the theoretical results on the error bounds presented in Section 4.2. Similar trends are observed for comparison of the non-RL NGD and SSP algorithms. In Figure 3e, where linkage error tends to zero, the post-RL versions of the three estimators approach the corresponding non-RL versions. NGD and SSP methods have strictly larger error than OLS due to the cost of privacy.

Figures 3b, 3d and 3f illustrate the empirical variances (EMP) against the theoretical variances (THR) of the proxy estimators given in Section 4.3. The theoretical variance of post-RL NGD closely aligns with the empirical variance at the chosen level of projection C. Recall that the theoretical variance would be exact when no projection is applied. Thus, with a lower level of projection on the gradient update, we anticipate it to be conservative. On the other hand, the theoretical variance of post-RL SSP approximates well with moderately large n and small σ. However, in scenarios with small n and/or large σ, our theoretical variance may underestimate the reality due to the approximation's reliance on a first-order Taylor expansion. Therefore, one can include higher-order terms for better approximation. In setting 3, where n and σ are fixed, as the linkage error vanishes, the variance reduces as a result of the smaller DP noise needed.

5.2. **Application to Synthetic Data.** Due to privacy concerns, pairs of datasets containing personal information, which serve as quasi-identifiers, are typically not made public. We instead synthesize from a pair of generated quasi-identifiers datasets and real data for regression, as in Chambers et al. (2021). For quasi-identifiers, we take advantage of the datasets generated by the Freely Extensible Biomedical Record Linkage (Febrl), which are available in the module `RecordLinkage` by Bruin (2022) in Python. The pair of datasets for linkage we use correspond to 5000 individuals. The domain indicator (state) is used for blocking. The record linkage is performed based on the Jaro-Winkler score (Jaro, 1989) or exact comparison on 6 quasi-identifiers (names, date of birth, address, etc.). The maximum score is 6 for pairs that have exact alignment. A threshold of 4 is chosen to link the records. For those left unlinked, we assign random links to ensure one-to-one linkage. A unique identifier is available in the dataset for verification purposes. The resulting linkage accuracy for the 9 blocks are $\boldsymbol{\gamma} = (0.880, 0.903, 0.918, 0.938, 0.905, 0.875, 0.898, 0.917, 0.898)$, making the overall accuracy 92.5%. We adopt the ELE model for Q and estimate it using $\boldsymbol{\gamma}$.

On the other hand, an anonymous dataset for regression comes from the Survey on Household Income and Wealth (SHIW) from the Bank of Italy (2012). The net disposable income and consumption are the explanatory variable X and the response $\boldsymbol{y}$, respectively. Since the SHIW dataset is larger, consisting of 8151 data points, we drop the outliers and randomly draw 5000 records and synthesize them with the Febrl dataset. Figure 4 depicts the setup of the synthesization process. Using the unique identifier from the Ferlb dataset, the regression variables $(X, \boldsymbol{y})$ are appended to the quasi-identifiers (Φ_A, Φ_B), resulting in two separate datasets: (Φ_A, X) and $(\Phi_B, \boldsymbol{y})$. Then, record linkage is performed by comparing Φ_A and Φ_B to output the linked data $(X, \boldsymbol{z})$ and the matrix Q.

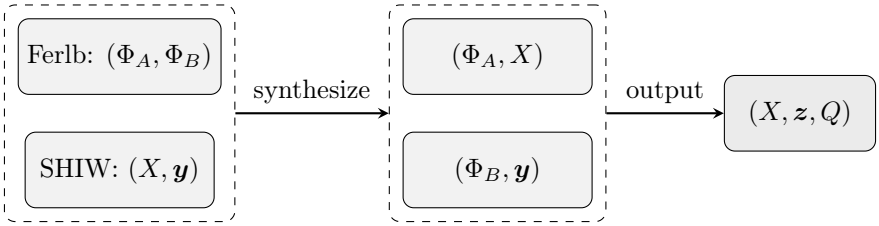


Figure 4. Synthesization. The Ferlb dataset provides quasi-identifiers (Φ_A, Φ_B), and the SHIW dataset provides regression variables $(X, \boldsymbol{y})$.

To apply the proposed DP algorithms to the synthesized dataset, we set the (hyper)parameters as follows. The privacy budget is given by $(\epsilon, \delta) = (1, 8.5 \times 10^{-5})$. The variance of the random error, σ^2, is estimated by the MSE. The upper bounds in Assumptions (A1)-(A3) are set as: $M = 1$, $c_0 = 1$, $c_x = \max(X) = 2.78$. In the NDG method, the projection level C is set to 1.2.

To illustrate the importance of propagating linkage uncertainty when conducting downstream regression, we also apply the non-RL version of NGD and SSP algorithms. We obtain the non-RL regression results by running post-RL NGD and post-RL SSP methods with M set to 0 and without converting X into W. This is equivalent to applying the non-RL methods discussed in Cai et al. (2021) and Sheffet (2017) to the linked set $(X, \boldsymbol{z})$ as if it were perfectly linked.

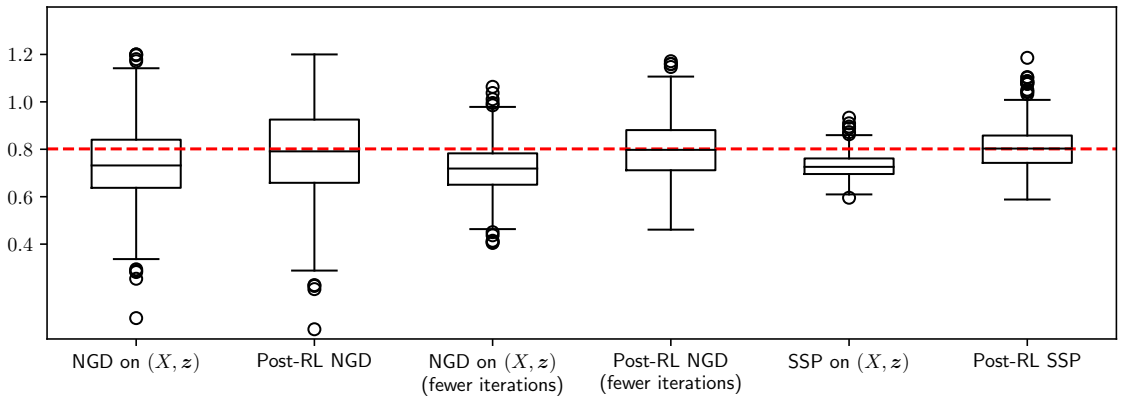


Figure 5. Boxplots of DP estimates based on 1000 repetitions with $(\epsilon, \delta) = (1, 8.5 \times 10^{-5})$. The red dashed line indicates the OLS estimate. The proposed post-RL algorithms are compared with the non-RL "NGD" and "SSP" methods applied to $(X, \boldsymbol{z})$ (i.e., without accounting present linkage errors). The third and fourth columns represent the two NGD methods running for $T = \lceil L^2 \ln(c_0^2 n)/3 \rceil$ iterations.

Figure 5 displays the boxplots of the estimates of each algorithm. For each algorithm, a total of 1000 repetitions are done in order to reflect the randomness of the injected noise for privacy purposes. The variables X and y are standardized before conducting simple linear regression. The OLS estimator on $(X, \boldsymbol{y})$ (dashed line) is plotted for comparison. As can be seen, the DP estimators by running (non-RL) NGD and SSP on $(X, \boldsymbol{z})$ directly are excessively biased as a consequence of ignoring linkage errors, even when the overall linkage accuracy is as high as 92.5%. Conversely, the results of post-RL NGD and post-RL SSP yields estimates centered around the OLS estimator but with higher variances, attributed to the cost of bias correction. Post-RL NGD is more flexible

due to hyperparameter tuning. Additionally, we run the NGD methods for fewer iterations with $T = \lceil L^2 \ln(c_0^2 n)/3 \rceil$, which is one-third of the value recommended by theory. We have found that this approach yields smaller variance while still producing accurate results in finite samples. Therefore, the theoretical number of iterations $T = \lceil L^2 \ln(c_0^2 n) \rceil$ may be conservative in some circumstances. Moderately reducing T may lead to better results.

6. Discussion

In this paper, we propose two differentially private algorithms for linear regression on a linked dataset that contains linkage errors, by leveraging the existing work on (1) linear regression after record linkage, and (2) differentially private linear regression. Figure 6 displays the connections among the related areas at a high level, including PPRL and SMPC mentioned in Sections 1 and 2.3. Our work is the first one to simultaneously consider the linkage uncertainty propagation and the privatization of the output. It also complements the area of PPRL where the main concern is the data leakage among different parties. However, we do not discuss how to link the records in the first place and thus the security issues of the linkage process are beyond our scope. Instead, we treat record linkage from a secondary perspective: we begin with linked data prepared by an external entity and we have limited information about the linkage quality.

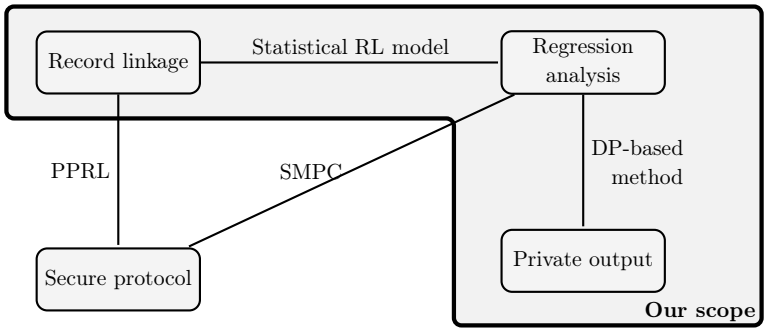


Figure 6. Diagram of related research areas. A secure protocol ensures no data is revealed to external parties during the linkage process.

Specifically, we propose two post-RL algorithms based on the noisy gradient descent and sufficient statistics perturbation methods from the DP literature. We provide privacy guarantees and finite-sample error bounds for these algorithms and discuss the variances of the private estimators. Our simulation studies and the application demonstrate the following: (1) the proposed estimators converge as the sample size increases; (2) post-RL linear regression incurs a higher cost than the non-RL counterpart in terms of the privacy-accuracy tradeoff; (3) The NGD method is flexible with hyperparameter tuning and can be applied to more general optimization problems; (4) SSP is specific to the least-squares problem, offering greater budget efficiency and more accurate results provided that the random error of the regression model is not too large.

There are different directions to extend our work. Note that there may be different scenarios of linking between the two datasets of the same set of entities. Assuming one-to-one linkage, as in our paper, is a canonical scenario. Although we do not explore it, we expect that our methods can be extended to other scenarios (e.g., one-to-many linkage) where Q still makes sense. Extra assumptions may be required when determining the relevant sensitivities for privacy purposes.

One can also consider record linkage from a primary perspective. In addition to the traditional Fellegi–Sunter model, Bayesian approaches and machine learning-based methods have gained popularity. The record linkage may take forms other than the matching probability matrix adopted here. Furthermore, when privacy concerns arise during the linkage process involving different parties, PPRL and SMPC protocols become essential. Tackling all the challenges depicted in Figure 6 simultaneously with a single efficient tool is of great practical use and significance. This interdisciplinary challenge requires expertise in both statistics and computer science.

Another important direction is exploring related statistical problems in the post-RL context, with or without privacy constraints. For example, confidence intervals and hypothesis testing are fundamental statistical inference tools. Other potential problems that interest statisticians include high-dimensional linear regression and ridge regression.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

Alabi, D., Smith, A., McMillan, A., Vadhan, S., & Sarathy, J. (2022). Differentially private simple linear regression. *arXiv:2007.05157*.

Alabi, D. G. (2022). The algorithmic foundations of private computational social science. *Harvard University Graduate School of Arts and Sciences*, Doctoral dissertation.

Apple. (2017). Learning with Privacy at Scale.

Bank of Italy. (2012). Survey on Household Income and Wealth.

Bassily, R., Smith, A. D., & Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 464–473.

Bernstein, G., & Sheldon, D. (2019). Differentially private bayesian linear regression. *Proceedings of the 33rd international conference on neural information processing systems* (pp. 525–535). Curran Associates Inc.

Binette, O., & Steorts, R. C. (2022). (almost) all of entity resolution. *Science Advances*, *8*(12), eabi8021. https://doi.org/10.1126/sciadv.abi8021

Bruin, J. (2022). Record Linkage Toolkit Documentation.

Bun, M., & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In M. Hirt & A. Smith (Eds.), *Theory of cryptography* (pp. 635–658). Springer Berlin Heidelberg.

Cai, T. T., Wang, Y., & Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, *49*(5), 2825–2850.

Chambers, R., Fabrizi, E., & Salvati, N. (2021). Small area estimation with linked data. *Journal of the Royal Statistical Society Series B, Royal Statistical Society*, *83(1)*, 78–107.

Chambers, R. (2009). Regression analysis of probability-linked data. *Technical report, Official Statistics Research, Statistics New Zealand*.

Chambers, R., Fabrizi, E., Ranalli, M. G., Salvati, N., & Wang, S. (2023). Robust regression using probabilistically linked data. *WIREs Computational Statistics*, *15*(2), e1596.

Chipperfield, J. (2020). Bootstrap inference using estimating equations and data that are linked with complex probabilistic algorithms. *Statistica Neerlandica*, *74*(2), 96–111.

Chipperfield, J. O., & Chambers, R. L. (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics*, *31*(3), 397–414. https://doi.org/10.1515/jos-2015-0024

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Publishing Company, Incorporated.

Christen, P., Ranbaduge, T., & Schnell, R. (2020). *Linking sensitive data: Methods and techniques for practical privacy-preserving information sharing*. Springer Cham. https://doi.org/10.1007/978-3-030-59706-1

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., & Stefanidis, K. (2020). An overview of end-to-end entity resolution for big data. *ACM Computing Surveys*, *53*(6). https://doi.org/10.1145/3418896

Dong, X. L., & Srivastava, D. (2015). Record linkage. *Big data integration* (pp. 63–106). Springer International Publishing.

Du, W., Han, Y. S., & Chen, S. (2004). Privacy-preserving multivariate statistical analysis: Linear regression and classification. In M. W. Berry, U. Dayal, C. Kamath, & D. B. Skillicorn (Eds.), *Proceedings of the fourth SIAM international conference on data mining* (pp. 222–233). SIAM.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., & Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In S. Vaudenay (Ed.), *Advances in cryptology - eurocrypt 2006* (pp. 486–503). Springer Berlin Heidelberg.

Dwork, C., & Nissim, K. (2004). Privacy-preserving datamining on vertically partitioned databases. In M. K. Franklin (Ed.), *Advances in cryptology - CRYPTO 2004, 24th annual international cryptologyconference* (pp. 528–544). Springer.

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

Dwork, C., Smith, A., Steinke, T., & Ullman, J. (2017). Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application (2017)*.

Dwork, C., Talwar, K., Thakurta, A., & Zhang, L. (2014). Analyze gauss: Optimal bounds for privacy-preserving principal component analysis. In D. B. Shmoys (Ed.), *Symposium on theory of computing, stoc 2014* (pp. 11–20). ACM. https://doi.org/10.1145/2591796.2591883

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Foulds, J., Geumlek, J., Welling, M., & Chaudhuri, K. (2016). On the theory and practice of privacy-preserving bayesian data analysis. *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 192–201.

Gascón, A., Schoppmann, P., Balle, B., Raykova, M., Doerner, J., Zahur, S., & Evans, D. (2017). Privacy-preserving distributed linear regression on high-dimensional data. *Proceedings on Privacy Enhancing Technologies*, *2017*(4), 345–364.

Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., & Kantarcioglu, M. (2021). Modern privacy-preserving record linkage techniques: An overview. *IEEE Transactions on Information Forensics and Security*, *16*, 4966–4987. https://doi.org/10.1109/TIFS.2021.3114026

Google. (2021). How we're helping developers with differential privacy.

Hall, R., & Fienberg, S. E. (2010). Privacy-preserving record linkage. In J. Domingo-Ferrer & E. Magkos (Eds.), *Privacy in statistical databases* (pp. 269–283). Springer Berlin Heidelberg.

Hall, R., Fienberg, S. E., & Nardi, Y. (2011). Secure multiple linear regression based on homomorphic encryption. *Journal of Official Statistics*, *27*, 669–691.

He, X., Machanavajjhala, A., Flynn, C., & Srivastava, D. (2017). Composing differential privacy and secure computation: A case study on scaling private record linkage. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1389–1406.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, *84*, 414–420.

Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E., & Malin, B. (2013). Efficient privacy-aware record integration. *Proceedings of the 16th International Conference on Extending Database Technology*, 167–178. https://doi.org/10.1145/2452376.2452398

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, *100*(469), 222–230. https://doi.org/10.1198/016214504000001277

Lin, S., Bun, M., Gaboardi, M., Kolaczyk, E. D., & Smith, A. (2023). Differentially private confidence intervals for proportions under stratified random sampling. *arXiv:2301.08324*.

Microsoft. (2020). Putting differential privacy into practice to use data responsibly.

Neter, J., Maynes, E. S., & Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, *60*, 1005–1027.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, *130*(3381), 954–959. https://doi.org/10.1126/science.130.3381.954

Ranbaduge, T., Vatsalan, D., & Ding, M. (2022). Privacy-preserving deep learning based record linkage. *CoRR:2211.02161*.

Rao, F.-Y., Cao, J., Bertino, E., & Kantarcioglu, M. (2019). Hybrid private record linkage: Separating differentially private synopses from matching records. *ACM Transactions on Privacy and Security*, *22*(3).

Sanil, A. P., Karr, A. F., Lin, X., & Reiter, J. P. (2004). Privacy preserving regression modelling via distributed computation. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 677–682.

Scheuren, F., & Winkler, W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, *19*, 39–58.

Sheffet, O. (2017). Differentially private ordinary least squares. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning, ICML 2017* (pp. 3105–3114). PMLR.

Steorts, R., Ventura, S., Sadinle, M., & Fienberg, S. (2014). A comparison of blocking methods for record linkage. *Domingo-Ferrer J. (eds) Privacy in Statistical Databases. PSD 2014.*, *8744*.

U.S. Census Bureau. (2021). Disclosure Avoidance for the 2020 Census: An Introduction.

U.S. Census Bureau. (2022). Annual Report of the Center for Statistical Research and Methodology.

Vu, D., & Slavkovic, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. *2009 IEEE International Conference on Data Mining Workshops*, 138–143. https://doi.org/10.1109/ICDMW.2009.52

Wang, Y.-X. (2018). Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. *Conference on Uncertainty in Artificial Intelligence (UAI)*, *49*.

Wang, Z., Ben-David, E., Diao, G., & Slawski, M. (2022). Regression with linked datasets subject to linkage error. *WIREs Computational Statistics*, *14*(4), e1570.

Zhang, L.-C., & Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society Series A, Royal Statistical Society*, *184(2)*, 522–547.