Differentially Private Linear Regression with Linked Data*

Shurong Lin
Department of Mathematics and Statistics, Boston University
Elliot Paquette
Department of Mathematics and Statistics, McGill University
Eric D. Kolaczyk
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 data sets 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.

*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 oftentimes scattered across different resources and no unique identifiers are available for perfect linkage. To conduct statistical modeling or inference on the integrated information, it is necessary to probabilistically link multiple data sets 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 Newcombe et al. (1959), Fellegi & Sunter (1969), Jaro (1989). Record linkage is still an active area of research and is used frequently in current practice. A prominent example is the U.S. decennial census (U.S. Census Bureau 2022). A recent review paper, Binette & Steorts (2022), provides a comprehensive summary. 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 data sets owned by different organizations are linked without revealing the data to one another (Hall & Fienberg 2010, Christen et al. 2020). PPRL, in turn, is associated with secure multiparty computation (SMPC) in that SMPC techniques are commonly used to solve PPRL problems (Kuzu et al. 2013, He et al. 2017, Rao et al. 2019). PPRL to date only engages in the private linkage process from a primary perspective, without concerning how the linkage errors 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 data set is subject to a variety of privacy attacks (Dinur & Nissim 2003, Narayanan & Shmatikov 2008, Dwork et al. 2017, Shokri 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, Microsoft 2020, Google 2021) 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.,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \mathcal{N}(0, \sigma^2 I_n) \tag{1}$$

but where X and $\boldsymbol{y}$ are observed in two separate data sets. 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 whole 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.

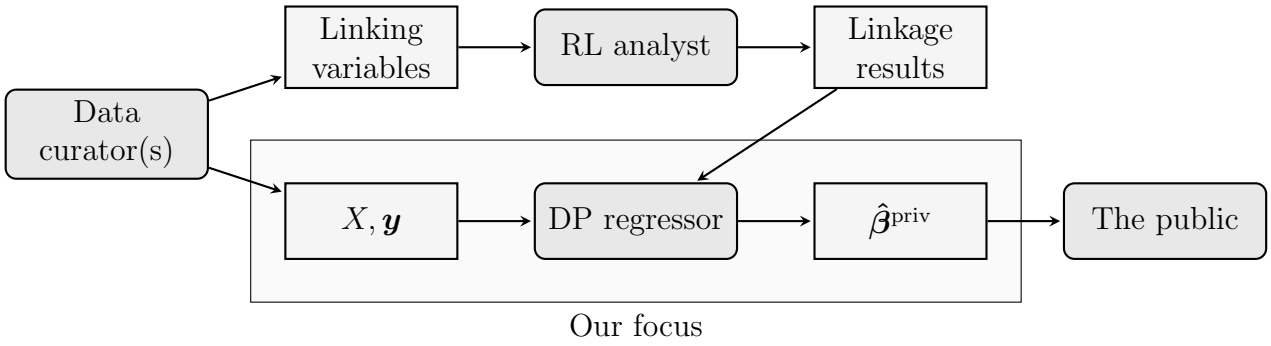


Figure 1: Pipeline of private regression with linked data. "Linkage variables" are quasi-identifiers like names, gender, etc. $\hat{\boldsymbol{\beta}}^{\text{priv}}$ denotes the differentially private (DP) estimator for the regression coefficients.

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 Lahiri & 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 (Algorithm 4.1) by Cai et al. (2021) and the "Analyze Gauss" algorithm for linear regression (Algorithm 2) by Sheffet (2017), respectively, to additionally handle the presence of linkage errors. These authors, in contrast, study differentially private linear regression with no prior linkage involved. If the data are linked beforehand, then 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). Our illustrative application later in the paper confirms this, where for example around 90% of the records are correctly linked and the estimates ignoring linkage errors end up with large biases.

Accompanying the estimators resulting from our algorithms, we provide finite-sample error bounds under typical regularity assumptions and record linkage schemes. In Theorems 4.4 and 4.5, we analyze the error solely resulting from differential privacy mechanisms for both algorithms, i.e., $\|\hat{\boldsymbol{\beta}}^{\mathrm{RL}} - \hat{\boldsymbol{\beta}}^{\mathrm{priv}}\|_2^2$ where $\hat{\boldsymbol{\beta}}^{\mathrm{RL}}$ is a non-private record linkage estimator proposed by Lahiri & Larsen (2005). The high-probability upper bound we obtain for $\|\hat{\boldsymbol{\beta}}^{\mathrm{RL}} - \hat{\boldsymbol{\beta}}^{\mathrm{priv}}\|_2^2$ in the gradient method shares the same order as the non-RL counterpart in Algorithm 4.1 by Cai et al. (2021), indicating that the error cost for privacy can be as low as the non-RL case ignoring a constant factor. According to Cai et al. (2021), their private estimator achieves a minimax lower bound up to logarithmic factors. Note that linkage errors do bring an additional and unavoidable error cost overall, which is reflected in Corollary 4.1. As for the sufficient statistics perturbation method, the error bound in Sheffet (2017) involves a random noise component from the DP mechanism whereas we offer a universal bound that is not DP noise-dependent. In addition, we have presented (approximate) theoretical variances for $\hat{\boldsymbol{\beta}}^{\mathrm{priv}}$ resulting from both algorithms. None of the existing work (Sheffet 2017, Wang 2018, Bernstein & Sheldon 2019, Cai et al. 2021, Alabi et al. 2022) on differentially private linear regression appear to have provided information about the variance of the private estimators.

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 data sets 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 $\boldsymbol{z}$:

$$\mathbb{P}(z_i = y_j) = q_{ij}, \quad i, j = 1, \ldots, n, \tag{2}$$

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 & 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), when inference is done after record

linkage based on $(X, \boldsymbol{z})$, Lahiri & Larsen (2005) proposed an unbiased estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{RL}} = (W^\top W)^{-1} W^\top \boldsymbol{z}, \tag{3}$$

where $W = QX$, with variance

$$\Sigma^{\mathrm{RL}} \stackrel{def}{=} \mathrm{Var}(\hat{\boldsymbol{\beta}}^{\mathrm{RL}}) = (W^\top W)^{-1} W^\top \Sigma_{\boldsymbol{z}} W (W^\top W)^{-1}, \tag{4}$$

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

Lemma 2.1 (Theorem A.1, Lahiri & Larsen (2005))**.** *Under the model described by (1) and (2), 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 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, Steorts et al. 2014, Christophides et al. 2020).

Exchangeable Linkage Errors (ELE) Model ELE model assumes homogeneous linkage accuracy and errors:

$$\begin{aligned} &\mathbb{P}(\text{correct linkage}) = q_{ii} = \gamma, \\ &\mathbb{P}(\text{incorrect linkage}) = q_{ij} = \frac{1-\gamma}{n-1} \text{ for } i \neq j. \end{aligned} \tag{5}$$

ELE model is adopted by Chambers (2009), Kim & Chambers (2012*a*,*b*), Chambers et al. (2021) for various estimation problems. Even though (5) 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 $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$ be two neighboring data sets of size n which only differ in one record. Such a relation is denoted by $\mathcal{D} \sim \mathcal{D}'$.

Definition 1 ((ϵ, δ)-DP, Dwork & 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 $\mathcal{D} \sim \mathcal{D}' \in \mathcal{X}^n$ and any $\mathcal{O} \subseteq \mathcal{R}$,*

$$\mathbb{P}(A(\mathcal{D}) \in \mathcal{O}) \leq e^{\epsilon} \cdot \mathbb{P}(A(\mathcal{D}') \in \mathcal{O}) + \delta. \tag{6}$$

The expression (6) controls the distance between the output distributions on two neighboring data sets through the privacy budget ϵ and δ. Intuitively, differential privacy ensures that $\mathcal{D}$ is not distinguishable from $\mathcal{D}'$ 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 (Composition, Dwork & Roth (2014)). *If f_1 is (ϵ_1, δ_1)-DP and f_2 is (ϵ_2, δ_2)-DP, then $f_1 \circ f_2$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.*

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

$$\Delta_2(f) = \max_{\mathcal{D} \sim \mathcal{D}' \in \mathcal{X}^n} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2. \tag{7}$$

The sensitivity of a function characterizes how much the output would change if one record in the data set 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.

Theorem 2.2 (Gaussian Mechanism, Dwork & Roth (2014)). *Let $\epsilon > 0$ and $\delta > 0$. For*

any algorithm f mapping a data set $\mathcal{D}$ to $\mathbb{R}^d$, the Gaussian Mechanism $A(\cdot)$ defined as

$$A(\mathcal{D}) = f(\mathcal{D}) + (u_1, \ldots, u_d)^\top, \tag{8}$$

where $u_1, \ldots, u_d \overset{i.i.d.}{\sim} \mathcal{N}(0, 2\ln(1.25/\delta)(\Delta_2(f)/\epsilon)^2)$, is (ϵ, δ)-DP.

2.3 Related Work

Linear regression with linked data is a fundamental statistical task that has been explored in various articles. Scheuren & Winkler (1993) initially considered the linkage model (2) for linear regression and proposed an estimator that is not generally unbiased. Later, Lahiri & Larsen (2005) introduced an exactly unbiased OLS-like estimator given in (3) with an expression for the variance, which outperformed the approach by Scheuren & Winkler (1993). Besides, Chambers (2009), Zhang & Tuoto (2021) offered a few other estimators. According to their simulation studies, some of the estimators provided at most similar but no noticeably better performance compared to the one proposed by Lahiri & Larsen (2005). Yet, Zhang & Tuoto (2021) relaxed the condition by not assuming that the probability of correct linkage, q_{ii} in the model (2), can be obtained or estimated. More broadly, other general regression approaches such as Bayesian methods and estimating equations with linked data have been discussed in Tancredi & Liseo (2015), Han & Lahiri (2019).

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 data sets 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 (He et al. 2017, 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 data set 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), Sanil et al. (2004), Hall et al. (2011), Gascón et al. (2017) discussed the implementations of privacy-preserving linear regression protocols that prevent data disclosure across organizations, whereas Dwork & 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 (3) 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), Cai et al. (2021), we tailor the noisy gradient method for the post-RL linear regression model for $(X, \boldsymbol{z})$ based on (1) and (2).

Let $\mathcal{L}_n(\boldsymbol{\beta}) \stackrel{def}{=} \frac{1}{2n}(\boldsymbol{z} - W\boldsymbol{\beta})^\top(\boldsymbol{z} - W\boldsymbol{\beta})$ 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 & Larsen (2005). Let $\Pi_R(\boldsymbol{r})$ denote the projection of $\boldsymbol{r} \in \mathbb{R}^s$ onto the ℓ_2 ball $\{\boldsymbol{r} \in \mathbb{R}^s : \|\boldsymbol{r}\| \leq R\}$. 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 $\boldsymbol{u}_t \sim \mathcal{N}(0, \omega^2 I_d)$ where $\omega = \dfrac{\eta T B \sqrt{2\ln(1.25T/\delta)}}{n\epsilon}$.

4: Compute

$$\boldsymbol{\beta}^{t+1} = \Pi_C(\boldsymbol{\beta}^t - \frac{\eta}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i^\top \boldsymbol{\beta}^t - \Pi_R(z_i))\boldsymbol{w}_i + \boldsymbol{u}_t). \tag{9}$$

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 the Gaussian Mechanism (Theorem 2.2) with the composition property (Proposition 2.1). By composition, it suffices to ensure that each iteration is $(\epsilon/T, \delta/T)$-DP for the final output $\hat{\boldsymbol{\beta}}^{\text{priv}}$ to satisfy (ϵ, δ)-DP. The scale of the Gaussian noise $\boldsymbol{u}_t$ at step t is determined by the privacy budget $(\epsilon/T, \delta/T)$, and the noise scale factor B associated with the sensitivity in the update function (9). 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 data sets $\mathcal{D} = (X, \boldsymbol{y})$ and $\mathcal{D}' = (X', \boldsymbol{y}')$ 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 data sets containing both linking variables and regression variables, denoted as $\mathcal{D} = (X, \Phi_X, \boldsymbol{y}, \Phi_{\boldsymbol{y}})$ and $\mathcal{D}' = (X', \Phi_{X'}, \boldsymbol{y}', \Phi_{\boldsymbol{y}'})$, 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 Vu & Slavkovic (2009), Foulds et al. (2016), Wang (2018). For the non-RL OLS estimator $\hat{\boldsymbol{\beta}}^{\mathrm{RL}} = (X^\top X)^{-1} X\boldsymbol{y}$, 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 used by Sheffet (2017) for the OLS estimation, which

essentially employed the SSP approach.

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 $A^\top A \stackrel{def}{=} \begin{pmatrix} W^\top W & W^\top \boldsymbol{z} \\ \boldsymbol{z}^\top W & \boldsymbol{z}^\top \boldsymbol{z} \end{pmatrix}$ 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 data set $(X, \boldsymbol{z})$ 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 $\mathcal{N}(0, \omega^2)$.

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. *It is possible that the matrix $W^\top W + U$ is nearly singular. In such cases, we discard the estimate and recompute. This does not affect the performance of the algorithm since the probability of $W^\top W + U$ being nearly singular is going to zero exponentially fast as 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 $c_{\boldsymbol{x}} < \infty$ such that $\|\boldsymbol{x}\| \leq c_{\boldsymbol{x}}$.

(A2) Let Q and Q' be the matching probability matrices (MPMs) resulting from the neighboring data sets $\mathcal{D}$ and $\mathcal{D}'$ 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.

Under Assumptions (A1) and (A2), it follows that

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

$$B = Rc_{\boldsymbol{x}}(M + 4) + 2Cc_{\boldsymbol{x}}^2(M + 2). \tag{10}$$

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

$$B = Rc_x(M+4) + \max\{2c_x^2(M+2), 2R^2\}. \tag{11}$$

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 linkage 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 linkage variables in the linkage model we consider. Specifically, we assume that two MPMs produced by two neighboring data sets do not differ much in terms of the entry-wise 1 norm. This assumption characterizes a *bounded linkage model*.

The rationality of (A2) is supported by typical schemes on the structures of MPM, as reviewed in Section 2.1.1. For example, with the blocking scheme, the size of each block is manageably small. When the linkage variable of one record is altered, the fluctuation of the MPM is limited to at most two blocks. Additionally, with the exchangeable linkage errors (ELE) model on each block, the change in one record is highly unlikely to significantly alter the linkage accuracy. A robust record linkage approach should not produce two considerably different MPMs from two neighboring data sets. Therefore, it is realistic to assume $\sup_{Q\sim Q'} \|Q - Q'\|_1 = O(1)$.

The proofs of Theorem 4.1 revolve around calculating the sensitivity of the target function in each algorithm. Besides the upper bounds $c_{\boldsymbol{x}}$ 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 $\|\boldsymbol{\beta}\| \leq c_0$ for some constant $0 < c_0 < \infty$.

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

$$0 < \frac{1}{L} < d\lambda_{\min}\left(\frac{W^\top W}{n}\right) \leq d\lambda_{\max}\left(\frac{W^\top W}{n}\right) < L \tag{12}$$

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. For the record linkage procedure to work moderately well, the MPM is sufficiently close to an identity matrix, and thus a reasonably larger L can be chosen to satisfy both (A4) and (A4'). Therefore, for convenience, we consider (A4) and (A4') 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,*

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{RL} - \boldsymbol{\beta}\|^2 = \operatorname{tr}(\Sigma^{RL}), \tag{13}$$

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 $\hat{\boldsymbol{\beta}}^{\mathrm{RL}}$ in (13) takes the reduced form $\sigma^2 \operatorname{tr}(X^\top X)^{-1}$ which is exactly the lower bound obtained by $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$. Then, by Lemma 4.2, we know that $\mathbb{E}\|\hat{\boldsymbol{\beta}}^{\mathrm{RL}} - \boldsymbol{\beta}\|^2$ 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 $\operatorname{tr}(\Sigma^{\mathrm{RL}})$ behaves in general.

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

Theorem 4.4 (Post-RL NGD)**.** *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* $B = R c_{\boldsymbol{x}}(4 + M) + 2 c_0 c_{\boldsymbol{x}}^2 (2 + M)$*;*

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

$$\|\hat{\boldsymbol{\beta}}^{priv} - \hat{\boldsymbol{\beta}}^{RL}\|^2 = O\left(\frac{1}{n} + \frac{\sigma^2 d^3 \ln^3 n \ln(1/\delta)}{n^2 \epsilon^2}\right). \tag{14}$$

Theorem 4.5 (Post-RL SSP)**.** *In Algorithm 2, set*

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

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

$$\|\hat{\boldsymbol{\beta}}^{priv} - \hat{\boldsymbol{\beta}}^{RL}\|^2 = O\left(\frac{\sigma^4 d^3 \ln^2 n \ln(1/\delta)}{n^2\epsilon^2}\right). \tag{15}$$

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). 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 (14). The first error term $\frac{1}{n}$ 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 $O\left(\frac{1}{n}\right)$, which is the same order as $\mathbb{E}\|\hat{\boldsymbol{\beta}}^{\text{OLS}} - \boldsymbol{\beta}\|^2$. However, more iteration results in a larger error, because the budget is $(\epsilon/T, \delta/T)$ for each iteration. As T increases, random noise with a larger scale is added at each iteration. This tradeoff should be handled with care. In practice, a smaller number of iterations may be favored for the tradeoff, especially when n is not sufficiently large.

For the SSP algorithm, the convergence rate in (15) 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)

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{priv} - \boldsymbol{\beta}\|^2 = O\left(\mathrm{tr}(\Sigma^{RL}) + \frac{\sigma^2 d^3 \ln^3 n \ln(1/\delta)}{n^2\epsilon^2}\right) \tag{16}$$

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

(ii) (Post-RL SSP)

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{priv} - \boldsymbol{\beta}\|^2 = O\left(\mathrm{tr}(\Sigma^{RL}) + \frac{\sigma^4 d^3 \ln^2 n \ln(1/\delta)}{n^2\epsilon^2}\right) \tag{17}$$

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

4.3 Variances

To the best of our knowledge, the majority of the work on differential privacy mainly focuses on the algorithm design and the discussion of the privacy-accuracy tradeoff. The variance of the differentially private estimator is barely discussed. Some 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 $\hat{\boldsymbol{\beta}}^{\text{priv}}$. To address this issue, we investigate the variance of $\hat{\boldsymbol{\beta}}^{\text{priv}}$ 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 $\hat{\boldsymbol{\beta}}^{\text{priv}}$.

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

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \frac{\eta}{n} W^\top (W\boldsymbol{\beta}^t - \boldsymbol{z}) + \boldsymbol{u}_t, \tag{18}$$

then the variance of the Tth iterate is given by

$$\Sigma = \sum_{t=1}^{T} (I_d - A)^{t-1} \cdot B^\top \Sigma_{\boldsymbol{z}} B \cdot \sum_{t=1}^{T} (I_d - A)^{t-1} + \omega^2 \sum_{t=1}^{T} (I_d - A)^{2t-2}, \tag{19}$$

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 (4).

The estimator in Algorithm 1 is a projected variant of (18). The use of projection with level C on $\boldsymbol{\beta}^{t+1}$ in (9) impedes the exact analysis of variance for $\hat{\boldsymbol{\beta}}^{\text{priv}}$. Instead, we provide the variance in (19) for the non-projected estimator as a conservative variance for $\hat{\boldsymbol{\beta}}^{\text{priv}}$. 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 (19) 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* $\hat{\boldsymbol{\beta}}' \xrightarrow{p} \hat{\boldsymbol{\beta}}^{priv}$ *as* $\omega \to 0$*. The variance of* $\hat{\boldsymbol{\beta}}'$ *is given by*

$$\Sigma = \Sigma^{RL} + \omega^2 (W^\top W)^{-1}(I_d + \Sigma_0 + \Sigma_1 + \Sigma_2)(W^\top W)^{-1}, \tag{20}$$

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 data set 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 & Larsen (2005) are included as benchmarks. The private, non-RL counterpart methods are also performed in the absence of linkage errors for comparison.

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). The parameters are set as follows:

- regression model: $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \text{uniform}(-1, 1)$, 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 1 to 2.5.
- ELE linkage model: blockwise linkage accuracy $\gamma_i \in \text{uniform}(0.6, 0.9)$ which characterizes Q, block size $n_i = 25$, $M = 1$ in Assumption (A2).
- privacy budget: $\epsilon = 1$, $\delta = 1/n^{1.1}$.

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 2 displays the ℓ_2 relative error and both empirical and theoretical variances for the two settings. For NGD methods, $T = \lceil L^2 \ln(c_0^2 n)/5 \rceil$ is used (i.e., one-fifth of the value recommended by the theory), which we have found yields better accuracy in finite samples.

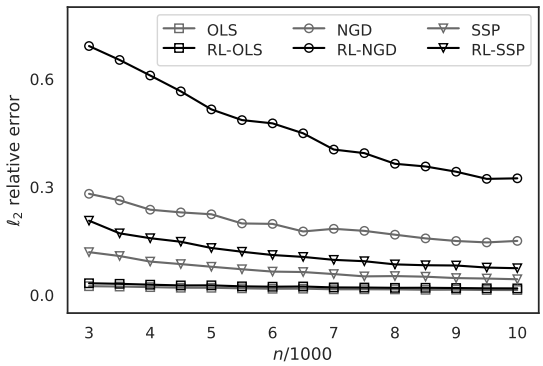
In setting 1, where σ is fixed at 1, Figure 2a shows the errors of all methods decrease with a growing sample size. Due to the linkage errors, the post-RL methods, including $\hat{\boldsymbol{\beta}}^{\mathrm{RL}}$ and our two algorithms (denoted as RL-OLS, RL-NGD, and RL-SSP in the figures)

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 $\sigma = 1$, SSP methods outperform NGD methods in terms of both accuracy and variance. However, as σ increases, NGD algorithms eventually perform better, as seen in Figure 2c 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.

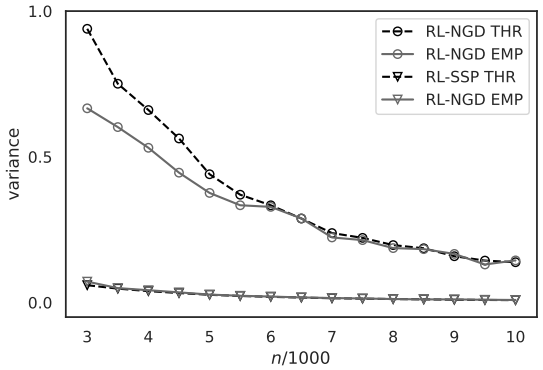
Figures 2b and 2d illustrate the empirical variances (EMP) against the theoretical variances (THR) of the proxy estimators given in Section 4.3. The RL-NGD theoretical variance gives reasonable estimations overall, but it becomes conservative when n is too small or σ is too large. The RL-SSP theoretical variance approximates well unless σ is too large (which can be mitigated by increasing n). The overestimation or underestimation of the variances using the proxy estimators is mainly due to the relatively large privacy noise resulting from small n or large σ.

5.2 Application to Synthetic Data

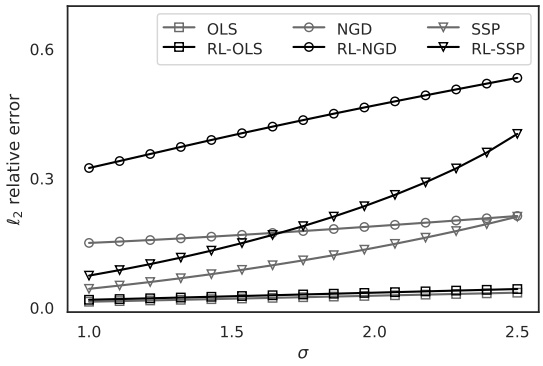
Since there is no single public data set that contains both record linkage variables and regression variables, we instead synthesize from two independent data sets, as in Chambers et al. (2021). For record linkage, we take advantage of the data sets generated by the Freely Extensible Biomedical Record Linkage (Febrl), which are available in the module `RecordLinkage` by Bruin (2022) in Python. The pair of data sets 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. The resulting linkage ac-



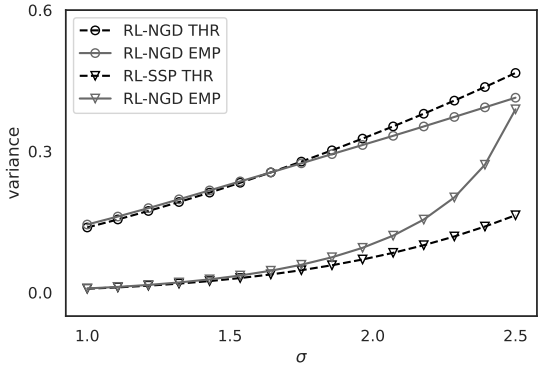
(a) Setting 1: $\sigma = 1$



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



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



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

Figure 2: Average ℓ_2-error and variance (theoretical and empirical) against n and σ, respectively.

curacy for the 9 blocks are $\boldsymbol{\gamma} = (0.880, 0.903, 0.918, 0.938, 0.905, 0.875, 0.898, 0.917, 0.898)$. We adopt the ELE model for Q and estimate it using $\boldsymbol{\gamma}$.

On the other hand, an anonymous data set 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 data set is larger, consisting of 8151 data points, we drop the outliers and randomly draw 5000 records and synthesize them with the Febrl data set. The variables X and y are standardized before conducting simple linear regression. The OLS estimate on $(X, \boldsymbol{y})$ is 0.805.

We apply the proposed differentially private algorithms to the synthesized data set with the privacy budget $(\epsilon, \delta) = (1, 8.5 \times 10^{-5})$. 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 2. For the SSP algorithm, a threshold of 20 is used to discard ineligible estimates, and new estimates are recomputed (Remark 3.1).

To illustrate that it is critical to propagate linkage uncertainty when conducting downstream regression, the non-RL counterparts of NGD and SSP algorithms are applied to the linked data set $(X, \boldsymbol{z})$ directly. For each algorithm, a total of 1000 repetitions are done in order to reflect the randomness of the noise injected for privacy. Figure 3 displays the boxplots of the estimates of each algorithm. NGD on $(X, \boldsymbol{z})$ and SSP on $(X, \boldsymbol{z})$ are the methods adapted from Cai et al. (2021), Sheffet (2017), respectively. NGD on $(W, \boldsymbol{z})$ and SSP on $(W, \boldsymbol{z})$ are the two algorithms we propose for post-RL regression. For the NGD algorithm, considering the default $c_x = 2.78$ leads to a large noise scale and is harmful to the accuracy, we also truncate X such that $c_x = 1.5$ and rerun the NGD algorithm. Around 10% of data are truncated and the new data set is denoted by $(X^{\mathrm{TR}}, \boldsymbol{z})$ or $(W^{\mathrm{TR}}, \boldsymbol{z})$. The OLS estimator on $(X, \boldsymbol{y})$ (dashed line) is plotted for comparison. As can be seen, the DP estimators on $(X, \boldsymbol{z})$ ignoring linkage errors are noticeably biased. The post-RL NGD and SSP algorithms correct the bias through transforming the design matrix $W = QX$, thus the results of NGD on $(W^{\mathrm{TR}}, \boldsymbol{z})$ and SSP on $(W, \boldsymbol{z})$ yields estimates centered around the OLS estimator, with higher variances due to bias correction.

6 Discussion

In this paper, we propose two differentially private algorithms for linear regression on a linked data set that contains linkage errors, by leveraging the existing work on (1) linear regression after record linkage, and (2) differentially private linear regression. Figure 4 displays the connections among the related areas at a high level, including PPRL and

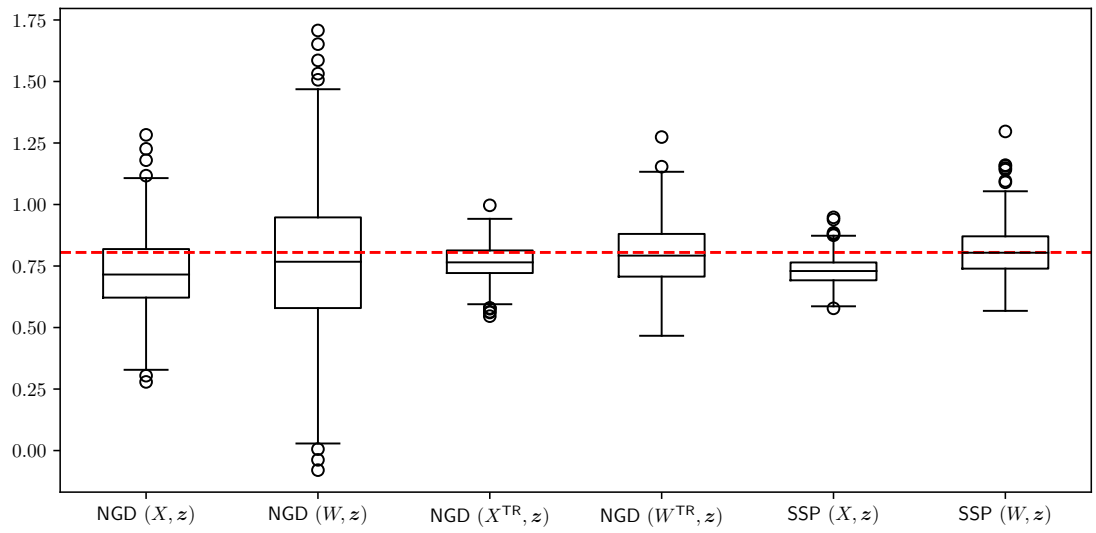


Figure 3: Boxplots of DP estimates based on 1000 repetitions with $(\epsilon, \delta) = (1, 8.5 \times 10^{-5})$. The red line indicates the OLS estimate. The algorithms run on $(W, \boldsymbol{z})$ or $(W^{\mathrm{TR}}, \boldsymbol{z})$ are post-RL methods, while those run on $(X, \boldsymbol{z})$ or $(X^{\mathrm{TR}}, \boldsymbol{z})$ are non-RL methods.

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.

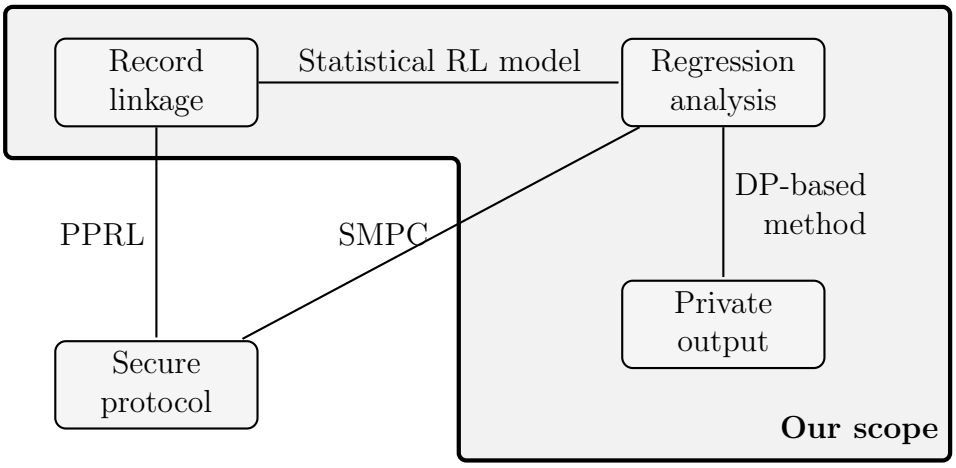


Figure 4: Diagram of related research areas. A secure protocol ensures no data is revealed to external parties during the linkage process.

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 non-RL linear regression in terms of the privacy-accuracy tradeoff; (3) the NGD method is more robust than the SSP method when dealing with input disturbances; and (4) SSP is more budget-efficient and yields more accurate results given the random error of regression model is not too large.

There are different directions to extend our work. One can 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 4 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.

Regarding differential privacy, one can explore alternative notions of DP. For example, zero-concentrated DP (Bun & Steinke 2016) is more efficient for basic composition. With the NGD method proposed here, the noise required for privacy purposes can be shown to scale up with $\sqrt{T}$ (where T is the number of iterations) instead of T. This supports the

scaling of T used in our simulations. Different variants of gradient descent are interesting to study, e.g., DP-SGD (Ponomareva et al. 2023). Additionally, introducing heterogeneous noise may prove helpful in achieving a better balance between accuracy and privacy tradeoffs.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. & Zhang, L. (2016), Deep learning with differential privacy, *in* "Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security", CCS '16, Association for Computing Machinery, p. 308–318.

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

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* pp. 464–473.

Bernstein, G. & Sheldon, D. (2019), *Differentially Private Bayesian Linear Regression*, Curran Associates Inc., p. 525–535.

Binette, O. & Steorts, R. C. (2022), "(almost) all of entity resolution", *Science Advances* **8**(12), eabi8021.

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

Bun, M. & Steinke, T. (2016), "Concentrated differential privacy: Simplifications, extensions, and lower bounds", *arXiv:1605.02065* .

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

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.

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.

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.

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).

Dinur, I. & Nissim, K. (2003), Revealing information while preserving privacy, *in* "Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems", PODS '03, Association for Computing Machinery, p. 202–210.

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

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", SIAM, pp. 222–233.

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", Springer Berlin Heidelberg, pp. 486–503.

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", Vol. 3152 of *Lecture Notes in Computer Science*, Springer, pp. 528–544.

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", ACM, pp. 11–20.

Fellegi, I. P. & Sunter, A. B. (1969), "A theory for record linkage", *Journal of the American Statistical Association* **64**(328), 1183–1210.

Foulds, J., Geumlek, J., Welling, M. & Chaudhuri, K. (2016), On the theory and practice of privacy-preserving bayesian data analysis, *in* "Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence", UAI'16, AUAI Press, p. 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.

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", Springer Berlin Heidelberg, pp. 269–283.

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

Han, Y. & Lahiri, P. (2019), "Statistical Analysis with Linked Data", *International Statistical Review* **87**(S1), 139–157.

He, X., Machanavajjhala, A., Flynn, C. & Srivastava, D. (2017), Composing differential privacy and secure computation: A case study on scaling private record linkage, *in* "Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security", CCS '17, Association for Computing Machinery, p. 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.

Kim, G. & Chambers, R. (2012*a*), "Regression analysis under incomplete linkage", *Computational Statistics & Data Analysis* **56**(9), 2756–2770.

Kim, G. & Chambers, R. (2012*b*), "Regression analysis under probabilistic multi-linkage", *Statistica Neerlandica* **66**, 64–79.

Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E. & Malin, B. (2013), Efficient privacy-aware record integration, *in* "Proceedings of the 16th International Conference on Extending Database Technology", EDBT '13, Association for Computing Machinery, p. 167–178.

Lahiri, P. & Larsen, M. D. (2005), "Regression analysis with linked data", *Journal of the American Statistical Association* **100**(469), 222–230.

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".

Narayanan, A. & Shmatikov, V. (2008), Robust de-anonymization of large sparse datasets, *in* "2008 IEEE Symposium on Security and Privacy (sp 2008)", pp. 111–125.

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.

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H., Vassilvitskii, S., Chien, S. & Thakurta, A. (2023), "How to dp-fy ml: A practical guide to machine learning with differential privacy".

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, *in* "Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", KDD '04, Association for Computing Machinery, p. 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", Vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 3105–3114.

Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017), Membership inference attacks against machine learning models, *in* "2017 IEEE Symposium on Security and Privacy (SP)", pp. 3–18.

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

Tancredi, A. & Liseo, B. (2015), "Regression analysis with linked data: problems and possible solutions", *Statistica* **75**(1), 19–35.

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, *in* "2009 IEEE International Conference on Data Mining Workshops", pp. 138–143.

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

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