Journal of Machine Learning Research 24 (2023) 1-50 Submitted 01/23; Revised 07/23; Published 11/23

Differentially Private Hypothesis Testing for Linear Regression

Daniel G. Alabi ALABID@CS.COLUMBIA.EDU
Data Science Institute
Columbia University
New York, NY 10027, USA

Salil P. Vadhan SALIL_VADHAN@HARVARD.EDU
John A. Paulson School of Engineering & Applied Sciences
Harvard University
Allston, MA 02134, USA

Editor: Moritz Hardt

Abstract

In this work, we design differentially private hypothesis tests for the following problems in the multivariate linear regression model: testing a linear relationship and testing for the presence of mixtures. The majority of our hypothesis tests are based on differentially private versions of the F-statistic for the multivariate linear regression model framework. We also present other differentially private tests—not based on the F-statistic— for these problems. We show that the differentially private F-statistic converges to the asymptotic distribution of its non-private counterpart. As a corollary, the statistical power of the differentially private F-statistic converges to the statistical power of the non-private F-statistic. Through a suite of Monte Carlo based experiments, we show that our tests achieve desired significance levels and have a high power that approaches the power of the non-private tests as we increase sample sizes or the privacy-loss parameter. We also show when our tests outperform existing methods in the literature.

Keywords: differential privacy, linear regression, robust statistics, small-area analysis

1. Introduction

Linear regression is one of the most fundamental statistical techniques available to social scientists and economists (especially econometricians) (Stock and Watson, 2011; Chetty et al., 2019). One of the goals of performing regression analysis is for use in decision-making via point estimation (i.e., getting a single predicted value for the dependent variable). To increase the confidence of decision-makers and analysts in such estimates, it is often important to also release accompanying uncertainty estimates for the point estimates (Chetty and Friedman, 2019; Chetty et al., 2018; Andrews et al., 2019).

In this work, we aim to provide differentially private linear regression *uncertainty quantification* via the use of hypothesis tests. Given the realistic possibility of reconstruction, membership, and inference attacks (Sweeney, 1997; Dwork et al., 2017), we can rely on Differential Privacy (DP), a rigorous approach to quantifying privacy loss (Dwork et al., 2006b,a). The task of DP linear regression is to, given datapoints $\{(x_i, y_i)\}_{i=1}^n$, [1] estimate point or uncertainty estimates for linear regression while satisfying differential privacy. The majority of our tests will rely on generalized likelihood ratio test F-statistics.

In previous works (Sheffet, 2017; Wang, 2018; Cai et al., 2021; Alabi et al., 2022), differentially private estimators for linear regression are explored and key factors (such as sample size and variance of the independent variable) that affect the accuracy of these estimators are identified. The focus of these previous works is for point estimate prediction. The predictive accuracy of such estimators can be measured in terms of a confidence bound or mean-squared error. We continue the study of the utility of such estimators for use in uncertainty quantification via hypothesis testing (Sheffet, 2017). (See Section 2 for background on hypothesis testing.)

Earlier work on uncertainty quantification for linear regression was done by Sheffet (2017), who constructed confidence intervals and hypothesis tests based on the t-test statistic, and can be used to test a linear relationship. The random projection routine in (Sheffet, 2017), based on the Johnson–Lindenstrauss (JL) transform, only starts to correctly reject the null hypothesis when the sample size is very large (or the variables have a large spread). This observation is also supported by the work of Couch et al. (2019). Furthermore, the random projection routine requires extra parameters (e.g., for specifying the dimensions of the random matrix). In our work, we use the F-statistic and our framework can be used to test mixture models, amongst other tests. We provide a general framework for DP tests based on the F-statistic. In addition, we will consider hypothesis testing for linear regression coefficients on both small and large datasets. For the mixture model tester, we additionally adapt and evaluate a nonparametric method, a DP analogue of the Kruskal-Wallis test due to Couch et al. (2019), which is especially suited for the small dataset regime. To the best of our knowledge, our tests are the first to differentially privately detect mixtures in linear regression models, with accompanying experimental validation.

1.1 Our Contributions

A short conference version of some of this work has appeared in (Alabi and Vadhan, 2022). We significantly extend their work, providing new non-parametric testing algorithms, with corresponding proofs and experimental analyses.

In this work, we show that for the problem of differentially private linear regression, we can perform hypothesis testing for two problems in the multivariate linear regression model:

1. **Testing a Linear Relationship**: is the slope of the linear model equal to some constant (e.g., slope is 0)?
2. **Testing for Mixtures**: does the population consist of one or more sub-populations with different regression coefficients?

We provide a differentially private analogue of the F-statistic which we, under the multivariate linear regression model, show converges in distribution to the asymptotic distribution of the F-statistic (Theorem 9). Furthermore, the DP regression coefficients converge, in probability, to the true coefficients (Lemma 18). In particular, in Lemma 18, we show

1. Where there exists $p \in \mathbb{N}$ such that for all $i \in [n]$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$.

a $1/\sqrt{n}$ statistical rate of convergence for the DP regression coefficients used for our hypothesis tests. This matches the optimal rate (Lei, 2011). We then use our DP F-statistic and parametric estimates to obtain DP hypothesis tests using a Monte Carlo parametric bootstrap, following Gaboardi et al. (2016). The Monte Carlo parametric bootstrap is used to ensure that our tests achieve a target significance level of α (i.e., data generated under the null hypothesis is rejected with probability α). We experimentally compare these tests to their non-private counterparts for univariate linear regression (i.e., one independent variable and one dependent variable). To the best of our knowledge, our tests are the first that use the F-statistic to perform tests on the problem of linear regression while ensuring privacy of the data subjects. In addition, our F-statistic based tests can be adapted to work on design matrices in any dimension. i.e., the design matrix can be cast in the form $X \in \mathbb{R}^{n \times p}$ for any integer $p \geq 2$ where n represents the number of individuals in the dataset and p represents the number of features per individual; we leave experimental evaluation of the multivariate case for future work.

Experimental evaluation of our hypothesis tests is done on:

1. **Synthetic Data**: We generate synthetic datasets with different distributions on the independent (or explanatory) variables. Specifically, we consider uniform, normal, and exponential distributions on the independent variables. We also vary the noise distribution of the dependent variable.

2. **Opportunity Insights (OI) Data**: We use a simulated version of the data used by the Opportunity Insights team (an economics research lab) to release the Opportunity Atlas tool, primarily used to predict social and economic mobility. We chose to use these datasets since they come from a real deployment of privacy-preserving statistics. [2] The census tract-level datasets from these states can have a very small number of datapoints.

3. **Bike Sharing Dataset**: We use a real-world dataset publicly available in the UCI machine learning repository. The dataset consists of daily and hourly counts (with other information such as seasonal and weather information) of bike rentals in the Capital bikeshare system in years 2011 and 2012.

Our experimental findings are as follows:

1. **Significance**: Across a variety of experimental settings, the significance is below the target significance level of 0.05. Thus, we have a high confidence that we will not falsely reject the null hypothesis, should the null hypothesis be true. Our tests are designed to be conservative in the sense that they err on the side of failing to reject the null (e.g., when the DP estimate of a variance is negative).

2. **Power**: Consistently, the power of our tests increase as we increase sample sizes (from hundreds up to tens of thousands) or as we relax the privacy parameter. The behavior of our DP tests tends toward that of the non-DP tests. The power of the DP OLS

2. See (Chetty et al., 2018, 2019) for a more detailed description of the use of the Opportunity Atlas tool in predicting social and economic mobility. (Chetty and Friedman, 2019; Alabi et al., 2022) evaluate privacy-protection methods on Opportunity Atlas data.

linear relationship tester increases as the slope of the model increases and as the noise in the dependent variable decreases. But when the DP estimate of the noise in the dependent variable is negative, the tests err on the side of failing to reject the null, leading to a lower power. The power of the mixture model tests also increases as the difference between the slopes in the two groups increases. And the more uneven the group sizes are, the lower the power.

3. **Alternative Method 1**: We compare our DP linear relationship tester, based on the F-statistic, to a test that builds on a DP parametric bootstrap method for confidence interval estimation due to Ferrando, Wang, and Sheldon (Ferrando et al., 2021). They prove that these intervals are consistent (in the asymptotic regime) and experimentally show that these intervals have good coverage. We can rely on such confidence intervals to decide to reject or fail to reject the null hypothesis. Such methods achieve the desired significance levels. However, we observe that the method is less powerful than the F-statistic approach. This behavior could be attributed to the differences in the bootstrap process: whereas we use estimates of sufficient statistics under the null in the bootstrap procedure, Ferrando et al. (Ferrando et al., 2021) uses the entirety of the sufficient statistics estimated for the parametric model.

4. **Alternative Method 2**: Inspired by the DP regression estimators of (Dwork and Lei, 2009; Alabi et al., 2022), we also show how to reduce linear relationship testing to the problem of testing if data is generated from a Bernoulli distribution with equal chance of outputting heads or tails. This problem has simple DP tests and has been solved optimally by Awan and Slavkovic (Awan and Slavkovic, 2020) for pure DP (whereas we use zCDP). The resulting linear relationship tester is nonparametric in that, in contrast to the F-statistic, it does not assume that the residuals are normally distributed. We find that this nonparametric test outperforms our DP F-statistic on smaller privacy-loss parameters or larger slopes, but otherwise the F-statistic performs better.

5. **Alternative Method 3**: For testing mixture models, we also give a reduction to testing whether two groups have the same median, which can be solved using the DP nonparametric method of Couch, Kazan, Shi, Bray, and Groce (Couch et al., 2019). We find that this nonparametric test has a higher power in the small dataset regime than the DP F-statistic method. As the dataset size increases or the difference in slopes between the groups increases, the gap closes. In addition, as the variance of the independent variable increases, the F-statistic method outperforms the nonparametric method.

1.2 Additional Related Work

1.2.1 Differentially Private Linear Regression

Sheffet (2017) considers hypothesis testing for ordinary least squares for a specific test: testing for a linear relationship under the assumption that the independent variable is drawn from a normal distribution. Wang (Wang, 2018) focuses on using adaptive algorithms for linear regression prediction.

M-estimators (Huber and Ronchetti, 2011), motivated by the field of robust statistics (Huber, 1964), are a simple class of statistical estimators that present a general approach to parametric inference. Dwork and Lei (Dwork and Lei, 2009; Lei, 2011) and Chaudhuri and Hsu (Chaudhuri and Hsu, 2012) present differentially private M-estimators with near-optimal statistical rates of convergence. Avella-Medina (Avella-Medina, 2020) generalizes the M-estimator approach to differentially private statistical inference using an empirical notion of influence functions to calibrate the Gaussian mechanism. Alabi, McMillan, Sarathy, Smith, and Vadhan (Alabi et al., 2022) proposed median-based estimators for linear regression and evaluated their performance for prediction. All of these previous works show connections between robust statistics, M-estimators, and differential privacy. The ordinary least squares estimator is a classical M-estimator for prediction. Other examples include sample quantiles and the maximum likelihood estimation (MLE) objective. However, for differentially private hypothesis testing, as we show in this work, the optimal test statistic for DP linear regression depends on statistical properties of the dataset (such as variance and sample size). We also present novel differentially private test statistics that converge in distribution to the asymptotic distribution of the F-statistic.

Bernstein and Sheldon (Bernstein and Sheldon, 2019) take a Bayesian approach to linear regression prediction and credible interval estimation. Through the Bayesian lens, there is also work on how to approximately bias-correct some DP estimators while providing some uncertainty estimates in terms of (private) standard errors (Evans et al., 2019). As a motivation for differentially private simple linear regression point and uncertainty estimation, Bleninger, Dreschsler, Ronning (Bleninger et al., 2010) show how an attacker could use background information to reveal sensitive attributes about data subjects used in a simple linear regression analysis.

1.2.2 General Differentially Private Hypothesis Testing

Gaboardi, Lim, Rogers, and Vadhan (Gaboardi et al., 2016) study hypothesis testing subject to differential privacy constraints. The tests they consider are: (1) *goodness-of-fit* tests on multinomial data to determine if data was drawn from a multinomial distribution with a certain probability vector, and (2) *independence* tests for checking whether two categorical random variables are independent of each other. Both tests use the chi-squared test statistic. Rogers and Kifer (Rogers and Kifer, 2017) develop new test statistics for differentially private hypothesis testing on categorical data while maintaining a given Type I error. Through the use of the subsample-and-aggregate framework, Barrientos et al. (Barrientos et al., 2017) compute univariate t-statistics by first partitioning the data into M disjoint subsets, estimating the statistic on each subset, truncating the statistic at some threshold a, and then adding noise from a Laplace distribution to the average of the truncated t-statistics. Our framework requires a clipping parameter (similar to a) but does not require any others (e.g., the number of subsets). As noted in that work, the parameter M plays a significant role in the performance—and tuning—of their tests while our tests require no such parameter tuning on partitioning of the data.

A subset of previous work (Task and Clifton, 2016; Campbell et al., 2018; Swanberg et al., 2019; Couch et al., 2019) focus on differentially private independence tests between a categorical and continuous variables. Some of these works produce nonparametric tests which require little or no distributional assumptions on the data generation process. Specifically, Couch, Kazan, Shi, Bray, and Groce (Couch et al., 2019) develop DP analogues of rank-based nonparametric tests such as Kruskal-Wallis and Mann-Whitney signed-rank tests. The Kruskal-Wallis test, for example, can be used to determine whether the medians of two or more groups are the same. We adapt their test to the setting of linear regression by using it to compare the distributions of slopes between the two groups. Wang, Lee, and Kifer (Wang et al., 2015) develop DP versions of likelihood ratio and chi-squared tests, showing a modified equivalence between chi-squared and likelihood ratio tests.

In the space of differentially private hypothesis testing, previous work introduce methods for differentially private identity and equivalence testing of discrete distributions (Acharya et al., 2019, 2018; Aliakbarpour et al., 2019, 2018). A differentially private version of the log-likelihood ratio test for the Neyman-Pearson lemma has also been shown to exist (Canonne et al., 2019). Furthermore, Awan and Slavkovic (Awan and Slavkovic, 2020) derive uniformly most powerful DP tests for simple hypotheses for binomial data. Suresh (Suresh, 2020) proposes a hypothesis test, which can be made to satisfy differential privacy, that is robust to distribution perturbations under Hellinger distance. Sheffet and Kairouz, Oh, Viswanath (Sheffet, 2018; Kairouz et al., 2016) also consider hypothesis testing, although in the local setting of differential privacy.

1.3 Organization of the Paper

In Section 2, we introduce some background notation and preliminaries in differential privacy, probability theory, and general hypothesis testing. In the non-private context (Section 3), we introduce further notation, formulation, and hypothesis tests for testing a linear relationship and testing for mixtures in data. Then we proceed to, in Section 4, present our differentially private tools and algorithms for hypothesis testing in the multivariate linear regression model. In Section 5, our main result is that the differentially private F-statistic converges to the asymptotic distribution of its non-private counterpart, the chi-squared distribution. We also show the $1/\sqrt{n}$ convergence rates for the DP sufficient statistics. In Section 6, we empirically evaluate the effectiveness of our hypothesis under different data distributions.

2. Preliminaries

2.1 Differential Privacy

For the definitions below, we say that two databases $\mathbf{x}$ and $\mathbf{x}'$ are neighboring, expressed notationally as $d(\mathbf{x}, \mathbf{x}') = 1$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$, if $\mathbf{x}$ differs from $\mathbf{x}'$ in exactly one row.

Definition 1 (Differential Privacy (Dwork et al., 2006b,a)) *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ be a (randomized) mechanism. For any $\epsilon \geq 0, \delta \in [0,1]$, we say $\mathcal{M}$ is (ϵ, δ)-**differentially private** if for all neighboring databases $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$, $d(\boldsymbol{x}, \boldsymbol{x}') = 1$ and every $S \subseteq \mathcal{R}$,*

$$\mathbb{P}[\mathcal{M}(\boldsymbol{x}) \in S] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(\boldsymbol{x}') \in S] + \delta.$$

If $\delta = 0$, then we say $\mathcal{M}$ is ϵ-DP, sometimes referred to as **pure** differential privacy. Typically, ϵ is a small constant (e.g., $\epsilon \in [0.1, 1]$) and $\delta \leq 1/\operatorname{poly}(n)$ is cryptographically small.

Definition 2 (Zero-Concentrated Differential Privacy (zCDP) (Bun and Steinke, 2016)) *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ be a (randomized) mechanism. For any neighboring databases $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$, $d(\boldsymbol{x}, \boldsymbol{x}') = 1$, we say $\mathcal{M}$ satisfies **ρ-zCDP** if for all $\alpha \in (1, \infty)$,*

$$D_\alpha(\mathcal{M}(\boldsymbol{x}) \| \mathcal{M}(\boldsymbol{x}')) \leq \rho \cdot \alpha,$$

where $D_\alpha(\mathcal{M}(\boldsymbol{x}) \| \mathcal{M}(\boldsymbol{x}'))$ is the Rényi divergence of order α between the distribution of $\mathcal{M}(\boldsymbol{x})$ and the distribution of $\mathcal{M}(\boldsymbol{x}')$. [3]

In this paper, we will primarily use ρ-zCDP as our definition of differential privacy, adding noise from a Gaussian distribution to ensure zCDP.

2.2 General Hypothesis Testing

The goal of hypothesis testing is to infer, based on data, which of two hypothesis, H_0 (the null hypothesis) or H_1 (the alternative hypothesis), should be rejected.

Let P_θ be a family of probability distributions parameterized by $\theta \in \Omega$. For some unknown parameter $\theta \in \Omega$, let $Z \sim P_\theta$ be the observed data. Then the two competing hypothesis are:

$$H_0 : \theta \in \Omega_0 \text{ vs. } H_1 : \theta \in \Omega_1,$$

where (Ω_0, Ω_1) form a partition of Ω.

A **test statistic** T is a random variable that is a function of the observed data $Z \sim P_\theta$. T can be used to decide whether to reject or fail to reject the null hypothesis. A **critical region** S is the set of values for the test statistic T (or correspondingly for the observed data) for which the null hypothesis will be rejected. It can be used to completely determine a test of H_0 versus H_1 as follows: We reject H_0 if $Y \in S$ and fail to reject H_0 if $Y \notin S$.

Sometimes, *external randomization* might help with choosing between hypothesis H_0 and H_1 (Edgington, 2011; Keener, 2010). By external randomness, we mean randomness not inherent in the sample or data collection process. In order to discriminate between hypothesis H_0 and H_1, we can define a notion of a critical function that can indicate the degree to which a test statistic is within a critical region. A critical function ϕ with range in $[0, 1]$ characterize randomized hypothesis tests. A nonrandomized test with critical region S can thus be specified as $\phi = 1_S$. Conversely, if $\phi(y)$ is always 0 or 1 for all y then the critical region is $S = \{y \ : \ \phi(y) = 1\}$ for this nonrandomized test. An advantage of allowing randomization (even without DP constraints) is that convex combinations of nonrandomized tests are not possible, but convex combinations of randomized tests are possible. i.e., if ϕ_1, ϕ_2 are critical functions and $t \in (0, 1)$, then $t\phi_1 + (1 - t)\phi_2$ is also a critical function so that the set of all critical functions form a convex set. Furthermore, nontrivial differentially private tests must be randomized.

For any $\theta \in \Omega$, the ideal test would tell us when $\theta \in \Omega_0$ and when $\theta \in \Omega_1$. This can be described by a **power function** $R(\cdot)$, which gives the chance of rejecting H_0 as a function of $\theta \in \Omega$:

$$R(\theta) = \mathbb{P}_\theta(Y \in S),$$

for any critical region S.

3. A related differential privacy notion, in terms of the Rényi divergence, is given in (Mironov, 2017).

A "perfect" hypothesis test would have $R(\theta) = 0$ for every $\theta \in \Omega_0$ and $R(\theta) = 1$ for every $\theta \in \Omega_1$. But this is generally impossible given only the "noisy" observed data $Z \sim P_\theta$.

A **significance level** α can be defined as

$$\alpha = \sup_{\theta \in \Omega_0} \mathbb{P}_\theta(Y \in S).$$

In other words, the level α is the worst chance of incorrectly rejecting H_0. Ideally, we want tests that have a small chance of error when H_0 should not be rejected. The **p-value** is the probability of finding, based on observed data, test statistics at least as extreme as when the null hypothesis holds. That is, if T is the test statistic function and t is the observed test statistic, then the (one-sided) p-value is $\mathbb{P}[T \geq t \mid H_0]$.

3. Hypothesis Testing for Linear Regression

In this section, we review the theory of (non-private) hypothesis testing in the multivariate linear regression model. We will consider hypothesis testing in the linear model

$$Y = X\beta + \mathbf{e},$$

where $X \in \mathbb{R}^{n \times p}$ is a matrix of known constants, $\beta \in \mathbb{R}^p$ is the parameter vector that determines the linear relationship between X and the dependent variable Y, and $\mathbf{e}$ is a random vector such that for all $i \in [n]$, $\mathbb{E}[e_i] = 0$, $\text{var}[e_i] = \sigma_e^2$. Furthermore, for all $i \neq j \in [n]$, $\text{cov}(e_i, e_j) = 0$.

Note that the simple linear regression model, $y_i = \beta_2 + \beta_1 \cdot x_i + e_i$ for scalars x_i, y_i and e_i $\forall i \in [n]$, can be cast as a linear model as follows: $X \in \mathbb{R}^{n \times 2}$ where

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix}. \tag{1}$$

We will consider the multivariate linear regression model: $Y \sim \mathcal{N}(X\beta, \sigma_e^2 I_{n \times n})$, where $I_{n \times n}$ is the $n \times n$ identity matrix. Let ω be an r-dimensional linear subspace of $\mathbb{R}^p$ and ω_0 be a q-dimensional linear subspace of ω such that $0 \leq q < r$. We will consider hypothesis tests of the form:

1. H_0: $\beta \in \omega_0$.
2. H_1: $\beta \in \omega \setminus \omega_0$.

Let $\hat{\beta}$ and $\hat{\beta}^N$ denote the least squares estimates under the alternative and null hypothesis respectively. In other words,

$$\hat{\beta}^N = \underset{z \in \omega_0}{\text{argmin}} \|Xz - Y\|^2, \quad \hat{\beta} = \underset{z \in \omega}{\text{argmin}} \|Xz - Y\|^2.$$

The **test statistic** (the F-statistic) we will use is equivalent to the generalized likelihood ratio test statistic

$$T = \left(\frac{n-r}{r-q}\right) \cdot \frac{\|Y - X\hat{\beta}^N\|^2 - \|Y - X\hat{\beta}\|^2}{\|Y - X\hat{\beta}\|^2} \tag{2}$$

$$= \left(\frac{n-r}{r-q}\right) \cdot \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2} \tag{3}$$

$$= \frac{1}{r-q} \cdot \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{S^2}, \tag{4}$$

where $S^2 = \|Y - X\hat{\beta}\|^2/(n-r)$. The vectors $Y - X\hat{\beta}$ and $X\hat{\beta} - X\hat{\beta}^N$ can be shown to be orthogonal, so that $\|Y - X\hat{\beta}^N\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\hat{\beta}^N\|^2$ by the Pythagorean theorem (Keener, 2010).

When $r - q = 1$, this test is *uniformly most powerful unbiased* and for $r - q > 1$, the test is most powerful amongst all tests that satisfy certain symmetry restrictions (Keener, 2010).

Theorem 3 *For every $n \in \mathbb{N}$ with $n > r$, let $X = X_n \in \mathbb{R}^{n \times p}$ be the design matrix. Under the multivariate linear regression model $Y = Y_n \sim \mathcal{N}(X_n\beta, \sigma_e^2 I_{n\times n})$,*

$$T = T_n \sim F_{r-q,n-r}(\eta_n^2), \quad \eta_n^2 = \frac{\|X_n\beta - X_n\beta^N\|^2}{\sigma_e^2},$$

where $F_{n,m}$ is the F-distribution with parameters n, m, $\beta^N = \mathbb{E}[\hat{\beta}^N]$, q is the dimension of ω_0, and r is the dimension of ω with $0 \le q < r$.

Furthermore,

1. $$\|Y_n - X_n\hat{\beta}\|^2 \sim \mathcal{X}_{n-r}^2\sigma_e^2, \quad \|X_n\hat{\beta} - X_n\hat{\beta}^N\|^2 \sim \mathcal{X}_{r-q}^2(\eta_n^2)\sigma_e^2.$$

2. *If there exists $\eta \in \mathbb{R}$ such that $\frac{\|X_n\beta - X_n\beta^N\|^2}{\sigma_e^2} \to \eta^2$, then*

$$T = T_n \sim F_{r-q,n-r}(\eta_n^2) \xrightarrow{D} \frac{\chi_{r-q}^2(\eta^2)}{r-q}.$$

3. *We have*

$$\frac{\|Y_n - X_n\hat{\beta}\|^2}{n-r} \xrightarrow{P} \sigma_e^2.$$

The values $\beta = \mathbb{E}[\hat{\beta}], \beta^N = \mathbb{E}[\hat{\beta}^N]$ are the expected values of our parameter estimates under the alternative and null hypotheses respectively.

A self-contained proof of Theorem 3 can be found in (Alabi and Vadhan, 2022).

Above, the **noncentral F-distribution** $F_{n,m}(\lambda)$, with parameters n, m and noncentrality parameter λ is the distribution of $\frac{\chi_n^2(\lambda)/n}{\chi_m^2/m}$, the ratio of two scaled chi-squared random variables. $\chi_K^2(\lambda)$ is a random variable distributed according to a chi-squared distribution with K degrees of freedom and noncentrality parameter λ. That is, $\chi_K^2(\lambda)$ is distributed as the squared length of a $\mathcal{N}(v, I_{K\times K})$ vector where $v \in \mathbb{R}^K$ has length λ. Also, $\chi_K^2 \sim \chi_K^2(0)$.

3.1 Testing a Linear Relationship in Simple Linear Regression Models

Consider the model: $y_i = \beta_2 + \beta_1 \cdot x_i + e_i$, where $e_i \sim \mathcal{N}(0, \sigma_e^2)$ are i.i.d. random variables and $x_1, \ldots, x_n$ are constants that form the following design matrix for our problem

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix}.$$

In this case, $\omega = \mathbb{R}^2$ and $\omega_0 = \{\beta \in \mathbb{R}^2 : \beta_1 = 0\}$. As a result, our hypothesis is:

1. H_0: $\beta_1 = 0$.
2. H_1: $\beta_1 \neq 0$.

Note that $r = p = 2$ and $q = 1$.

Furthermore, let

$$\beta = \begin{pmatrix} \beta_2 \\ \beta_1 \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_1 \end{pmatrix}.$$

We use $\hat{\beta}^N$ to refer to the estimate of β when the null hypothesis is true (i.e., $\beta_1 = 0$) and $\hat{\beta}$ be the estimate of β when the alternative hypothesis holds.

For the calculations below, let

1. $\mathbf{x} \stackrel{\text{def}}{=} (x_1, x_2, \ldots, x_n)^T$, $\mathbf{y} \stackrel{\text{def}}{=} (y_1, y_2, \ldots, y_n)^T$,
2. $\bar{x} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n y_i$,
3. $\overline{x^2} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n x_i^2$, $\overline{xy} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n x_i y_i$,
4. $\widehat{\sigma_{xy}^2} \stackrel{\text{def}}{=} \overline{xy} - \bar{x} \cdot \bar{y}$, and $\widehat{\sigma_x^2} \stackrel{\text{def}}{=} \overline{x^2} - \bar{x}^2$.

We can then obtain the sufficient statistics

$$X^T X = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\overline{x^2} \end{pmatrix}, \; X^T Y = \begin{pmatrix} n\bar{y} \\ n\overline{xy} \end{pmatrix}, \tag{5}$$

so that under the alternative hypothesis, the least squares estimate is

$$\hat{\beta} = \underset{\beta \in \omega}{\operatorname{argmin}} \|Y - X\beta\|^2 = (X^T X)^{-1} X^T Y,$$

assuming that $X^T X$ is invertible which happens iff $\mathbf{x}$ is not the constant vector (so that $\det(X^T X) = n^2\overline{x^2} - n^2\bar{x}^2 = n^2 \cdot \widehat{\sigma_x^2} > 0$). Assuming invertibility, we have

$$(X^T X)^{-1} = \frac{1}{n^2\overline{x^2} - n^2\bar{x}^2} \begin{pmatrix} n\overline{x^2} & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}. \tag{6}$$

Thus, the least squares estimate under the alternative hypothesis is

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= \frac{1}{n^2 \cdot \overline{x^2} - n^2 \cdot \bar{x}^2} \begin{pmatrix} n^2 \cdot \overline{x^2} \cdot \bar{y} - n^2 \cdot \bar{x} \cdot \overline{xy} \\ -n^2 \cdot \bar{x} \cdot \bar{y} + n^2 \cdot \overline{xy} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_1 \end{pmatrix},\end{aligned}$$

and further simplification results in the following slope and intercept estimates:

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\widehat{\sigma^2_{xy}}}{\widehat{\sigma^2_x}},$$

$$\hat{\beta}_2 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\bar{y} \cdot \overline{x^2} - \bar{x} \cdot \overline{xy}}{\widehat{\sigma^2_x}}.$$

The square of residuals is $\|Y - X\hat{\beta}\|^2$ and an (unbiased) estimate of σ_e^2 is $S^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-2}$.

Also, we can derive $\hat{\beta}^N$ as follows

$$\hat{\beta}^N = \underset{\beta \in \omega_0}{\operatorname{argmin}} \|Y - X\beta\|^2 = \begin{pmatrix} \bar{y} \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_2^N \\ 0 \end{pmatrix}$$

so that

$$\begin{aligned}\|X\hat{\beta} - X\hat{\beta}^N\|^2 &= \sum_{i=1}^{n} (\hat{\beta}_2 + \hat{\beta}_1 x_i - \hat{\beta}_2^N)^2 \\ &= \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \hat{\beta}_1^2 \cdot n \cdot \widehat{\sigma^2_x}.\end{aligned}$$

As a result, the test statistic T is

$$T = \left(\frac{n-r}{r-q}\right) \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2} = \frac{\hat{\beta}_1^2}{S^2} \cdot n \cdot \widehat{\sigma^2_x}.$$

3.2 Testing for Mixtures in Simple Linear Regression Models

The goal of testing mixtures is to detect the presence of sub-populations. Consider the model where $n = n_1 + n_2, n_1, n_2 > 0$, $\beta_1, \beta_2 \in \mathbb{R}$ with the following generation model:

- $y_i = \beta_1 \cdot x_i + e_i$ for $i \in [n_1]$.
- $y_i = \beta_2 \cdot x_i + e_i$ for $i \in \{n_1 + 1, \ldots, n\}$.

where $e_i \sim \mathcal{N}(0, \sigma_e^2)$ are i.i.d. random variables and $x_1, \ldots, x_n$ are constants that form the following design matrix for our problem

$$X = \begin{pmatrix} x_1 & 0 \\ \vdots & \vdots \\ x_{n_1} & 0 \\ 0 & x_{n_1+1} \\ \vdots & \vdots \\ 0 & x_n \end{pmatrix}.$$

Note that X is of full rank (except if all the x_i's are 0 either for all $i \in [n_1]$ or for all $i \in \{n_1+1, \dots, n\}$). Furthermore, $r = p = 2$.

In this case, $\omega = \mathbb{R}^2$ and $\omega_0 = \{\beta \in \mathbb{R}^2 : \beta_1 = \beta_2\}$. As a result, our hypothesis is:

1. H_0: $\beta_1 = \beta_2$.

2. H_1: $\beta_1 \neq \beta_2$.

Furthermore, let

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}.$$

We use $\hat{\beta}^N$ to refer to the estimate of β when the null hypothesis is true (i.e., $\beta_1 = \beta_2$) and $\hat{\beta}$ be the estimate of β when the alternative hypothesis holds.

For the calculations below, let $n_2 = n - n_1$ and

- $\overline{x^2}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^2, \overline{x^2}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n} x_i^2, \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$
- $\overline{xy}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i y_i, \overline{xy}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n} x_i y_i, \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$

We can then obtain

$$X^T X = \begin{pmatrix} n_1\overline{x^2}_1 & 0 \\ 0 & n_2\overline{x^2}_2 \end{pmatrix}, X^T Y = \begin{pmatrix} n_1\overline{xy}_1 \\ n_2\overline{xy}_2 \end{pmatrix},$$

so that, assuming $\overline{x^2}_1, \overline{x^2}_2 > 0$, we have

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{pmatrix} \overline{xy}_1/\overline{x^2}_1 \\ \overline{xy}_2/\overline{x^2}_2 \end{pmatrix}.$$

Furthermore,

$$\hat{\beta}^N = \begin{pmatrix} \overline{xy}/\overline{x^2} \\ \overline{xy}/\overline{x^2} \end{pmatrix},$$

since under the null hypothesis ($\beta_1 = \beta_2$), the design matrix "collapses" to

$$X_0 = \begin{pmatrix} x_1 \\ \vdots \\ x_{n_1} \\ x_{n_1+1} \\ \vdots \\ x_n \end{pmatrix},$$

so that $X_0^T X_0 = \sum_{i=1}^n x_i^2 = n\overline{x^2}$ and $X_0^T Y = \sum_{i=1}^n x_i y_i = n\overline{xy}$.

The squares of residuals is $\|Y - X\hat{\beta}\|^2$ and an (unbiased) estimate of σ_e^2 is $S^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-2}$.

Lemma 4

$$\|X\hat{\beta} - X\hat{\beta}^N\|^2 = \frac{n_1\overline{x^2}_1 n_2\overline{x^2}_2}{n\overline{x^2}}(\hat{\beta}_1 - \hat{\beta}_2)^2,$$

where X is the design matrix, $\hat{\beta}, \hat{\beta}^N$ are the least squares estimates under the alternative and null hypothesis, respectively.

A self-contained proof of Lemma 4 can be found in (Alabi and Vadhan, 2022).

By Lemma 4, our test statistic T is

$$T = \left(\frac{n-r}{r-q}\right)\frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2} = \frac{n_1\overline{x^2}_1 n_2\overline{x^2}_2}{S^2 n\overline{x^2}}(\hat{\beta}_1 - \hat{\beta}_2)^2.$$

4. Differentially Private Monte Carlo Tests

Since the private test statistic differs from the non-private version, we have to create new statistics to account for the level-α Monte Carlo differentially private testing. The majority of our tests will be based on DP sufficient statistics. In the statistics literature, a statistic is considered **sufficient**, with respect to a particular model, if it provides at least as much information for the value of an unknown parameter as any other statistic that can be calculated on a given sample (Keener, 2010).

Previous work (Sheffet, 2017; Wang, 2018; Alabi et al., 2022) perturb the sufficient statistics for ordinary least squares and use the result to compute a slope and intercept in a DP way. To add noise to ensure privacy, we typically have to truncate certain random variables. Let $A, B \in \mathbb{R}$ where $A \leq B$. We use $Y|_B^A$ to mean that the range of the random variable Y will be truncated to have an upper bound of A and a lower bound of B.

For all our DP OLS Monte Carlo tests that sample from a continuous Gaussian, we can instead use discrete variants (e.g., (Canonne et al., 2020)). The DP OLS Monte Carlo tests below use the zero-concentrated differential privacy definition (Bun and Steinke, 2016).

4.1 Monte Carlo Hypothesis Testing

We now proceed to discuss our general approach for designing a Level-α test for the task of linear regression estimation based on sufficient statistic perturbation. We rely on a subroutine `DPStats` that can produce DP statistics, given a dataset and privacy parameters, when testing. Algorithm 1 can then be specialized to test for the presence of a linear relationship and for mixture models.

To design a Monte Carlo hypothesis test, we follow a similar route to Gaboardi, Lim, Rogers, and Vadhan (Gaboardi et al., 2016). In Algorithm 1, we provide a framework to perform DP Monte Carlo tests using a parametric bootstrap based on a test statistic. Note that Algorithm 1 spends the privacy budget once, during the first call to `DPStats`, to compute DP sufficient statistics of the data. The remaining calls to `DPStats` are on already-privatized data and thus need not expend additional privacy budget. Let `DPStats` be a procedure that uses one or more statistics of X, Y to produce DP statistics that can be used to reject or fail to reject the null hypothesis. In this paper, `DPStats` will satisfy

Algorithm 1: Monte Carlo DP Test Framework.

Data: $X \in \mathbb{R}^{n \times p}; Y \in \mathbb{R}^n$
Input: n (dataset size); ρ (privacy-loss parameter); α (target significance); T (test statistic)
$(\tilde{\theta}_0, \tilde{\theta}_1) = \texttt{DPStats}(X, Y, n, \rho)$
if $\tilde{\theta}_0 = \tilde{\theta}_1 = \perp$ **then**
  return Fail to Reject the null

// non-DP test statistic applied to DP statistics
$\tilde{T} = \tilde{t} = T(\tilde{\theta}_1)$
Select $K > 1/\alpha$
for $k = 1 \ldots K$ **do**
  Initialize ${}^kX, {}^ky$
  $\forall i \in [n]$, ${}^kX_i, {}^ky_i \sim P_{\tilde{\theta}_0}$
  ${}^k\tilde{\theta}_0, {}^k\tilde{\theta}_1 = \texttt{DPStats}({}^kX, {}^ky, n, \rho)$
  if ${}^k\tilde{\theta}_0 = {}^k\tilde{\theta}_1 = \perp$ **then**
    Set t_k to ∞
  else
    Obtain t_k from $T({}^k\tilde{\theta}_1)$

Sort $t_1, \ldots, t_K$ into $t_{(1)} \leq \cdots \leq t_{(K)}$

Set $r = \lceil (K+1)(1-\alpha) \rceil$
if $\tilde{t} > t_{(r)}$ **then**
  return Reject the null
else
  return Fail to Reject the null

ρ-zCDP (Zero-Concentrated Differential Privacy). [4] T is the test statistic computation procedure. As done in (Gaboardi et al., 2016), for example, we will assume the dataset sizes are public information.

Let $T = T(\hat{\theta}_1)$ be the non-private test statistic procedure given $\hat{\theta}_1 = \hat{\theta}_1(X, Y)$. The goal is to compute $T(\tilde{\theta}_1)$ where $\tilde{\theta}_1$ is an approximation of $\hat{\theta}_1$. `DPStats` returns $\tilde{\theta}_0$ and $\tilde{\theta}_1$. If $\tilde{\theta}_0$ and $\tilde{\theta}_1$ is not $\perp$ ($\perp$ is returned whenever the perturbed statistics cannot be used to simulate the null distributions), then we use $\tilde{\theta}_1$ to compute the DP test statistic and $\tilde{\theta}_0$ to simulate the null. $P_{\tilde{\theta}_0}$ represents the distribution from which we will sample from to simulate the null distribution. When $(X, y) \sim P_{\tilde{\theta}_0}$ for $\theta_0 \in \Omega_0$ and we set $\tilde{\theta}_1 = \tilde{\theta}_1(X, Y)$ and sample $(X', y') \sim P_{\tilde{\theta}_0}$, then $\hat{\theta}_1((X', y'))$ has approximately the same distribution as $\hat{\theta}_1((X, y))$.

4. Gaussian noise addition (for privacy) was chosen because the noise in the dependent variable is also assumed to be Gaussian. The use of the Gaussian (or truncated Gaussian) distribution for both privacy and sampling error is a convenient choice as it could result in a clearer, more compatible, theoretical analysis.

4.2 Testing a Linear Relationship

We now discuss our F-statistic and Bernoulli testing approaches.

4.2.1 F-Statistic

For testing a linear relationship in simple linear regression models, recall that in the non-private case, we had

$$T(X, Y, \hat{\beta}, \hat{\beta}^N, n, r, q) = \left(\frac{n-r}{r-q}\right) \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2}.$$

Accordingly, we define and compute $\tilde{T}_L(X, Y, \hat{\beta}, \hat{\beta}^N, n, r, q, \rho, \Delta) = \tilde{t}$, a private estimate of $T(X, Y, \hat{\beta}, \hat{\beta}^N, n, r, q)$. In Algorithm 2, we give the full ρ-zCDP procedure for computing all necessary sufficient statistics to compute $\tilde{T}_L(X, Y, \hat{\beta}, \hat{\beta}^N, n, r, q, \rho, \Delta)$.

The DP estimate of S^2, $\widetilde{S^2}$, can be computed as $\widetilde{S^2} = \frac{\sum_{i=1}^n (y_i - \tilde{\beta}_2 - \tilde{\beta}_1 x_i)^2|_0^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho})}{n-r}$. Another equivalent way to compute $\widetilde{S^2}$ is to compute $\widetilde{y^2}$ privately and then, together with the other DP estimates, to compute $\widetilde{S^2}$. Note that under the null hypothesis, the DP estimate of S^2 is $\widetilde{S_0^2} = \frac{\sum_{i=1}^n (y_i - \tilde{\beta}_2)^2|_0^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho})}{n-r}$ which can also be equivalently computed by using $\tilde{\bar{y}}, \widetilde{y^2}, \tilde{\beta}_2$. Also, we return $(\tilde{\theta}_0, \tilde{\theta}_1) = (\perp, \perp)$ if the computed DP sufficient statistics cannot be used to simulate the null distribution.

Lemma 5 *For any $\rho, \Delta > 0$, Algorithm 2 satisfies ρ-zCDP.*

Proof This follows from Proposition 1.6 in (Bun and Steinke, 2016) (use of the Gaussian Mechanism). Next, we apply composition and post-processing (Lemmas 1.7 and 1.8 in (Bun and Steinke, 2016)). The computation of the following statistics is each done to satisfy $\rho/5$-zCDP: $\tilde{\bar{x}}, \tilde{\bar{y}}, \widetilde{\overline{x^2}}, \widetilde{\overline{xy}}, \widetilde{\overline{y^2}}$. $\tilde{\beta}_1, \tilde{\beta}_2, \widetilde{S^2}, \widetilde{S_0^2}$ are post-processing of the other DP releases.

As a result, the entire procedure satisfies ρ-zCDP. ■

Instantiating Algorithm 1 for the Linear Tester: If the procedure $\texttt{DPStats}_L$ returns $(\perp, \perp)$, then we fail to reject the null. Otherwise, we use the returned statistics $\tilde{\theta}_1 = (\tilde{\beta}_1, \tilde{\bar{x}}, \widetilde{\overline{x^2}}, \widetilde{S^2}, n)$ to create the test statistic $T(\tilde{\theta}_1) = \frac{\tilde{\beta}_1^2 \cdot n \cdot (\widetilde{\overline{x^2}} - \tilde{\bar{x}}^2)}{\widetilde{S^2}}$ and use $\tilde{\theta}_0 = (\tilde{\bar{y}}, \tilde{\bar{x}}, \widetilde{\overline{x^2}}, \widetilde{S_0^2}, n)$ to simulate the null distributions (to decide to reject or fail to reject the null hypothesis). $P_{\tilde{\theta}_0}$ is instantiated as a normal distribution and used to generate ${}^k x_i$ distributed as $\mathcal{N}(\tilde{\bar{x}}, (n\widetilde{\overline{x^2}} - n\tilde{\bar{x}}^2)/(n-1))$ and ${}^k y_i$ as $\tilde{\beta}_2 + e_i$, $e_i \sim \mathcal{N}(0, \widetilde{S_0^2})$ for all $i \in [n]$.

4.2.2 Bernoulli Tester

Next, we define an approach for testing a linear relationship via Bernoulli testing, inspired by the DP regression estimators of (Dwork and Lei, 2009; Alabi et al., 2022).

Under the null hypothesis (slope is $\beta_1 = 0$), observe that, under the multivariate linear regression model, if we pair datapoints $(x_i, y_i), (x_{i+1}, y_{i+1})$ and calculate the sign at the

---

Algorithm 2: ρ-zCDP procedure $\texttt{DPStats}_L$

Data: $X \in \mathbb{R}^{n\times 2}; Y \in \mathbb{R}^n$

Input: integer $n \geq 2$; $r, q \in \mathbb{N}$; $\rho > 0, \Delta > 0$

Set $\rho_0 = \rho/5$ and compute the following:

1. $\tilde{\bar{x}} = \frac{1}{n}\sum_{i=1}^n x_i|_{-\Delta}^{\Delta} + \mathcal{N}(0, \frac{2\Delta^2}{\rho_0 n^2})$.
2. $\tilde{\bar{y}} = \frac{1}{n}\sum_{i=1}^n y_i|_{-\Delta}^{\Delta} + \mathcal{N}(0, \frac{2\Delta^2}{\rho_0 n^2})$.
3. $\widetilde{\overline{x^2}} = \frac{1}{n}\sum_{i=1}^n x_i^2|_{0}^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho_0 n^2})$.
4. $\widetilde{\overline{xy}} = \frac{1}{n}\sum_{i=1}^n x_i y_i|_{-\Delta^2}^{\Delta^2} + \mathcal{N}(0, \frac{2\Delta^4}{\rho_0 n^2})$.
5. $\widetilde{\overline{y^2}} = \frac{1}{n}\sum_{i=1}^n y_i^2|_{0}^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho_0 n^2})$.
6. $$\tilde{\beta}_1 = \frac{\widetilde{\overline{xy}} - \tilde{\bar{x}}\tilde{\bar{y}}}{\widetilde{\overline{x^2}} - \tilde{\bar{x}}^2}, \quad \tilde{\beta}_2 = \frac{\tilde{\bar{y}} \cdot \widetilde{\overline{x^2}} - \tilde{\bar{x}} \cdot \widetilde{\overline{xy}}}{\widetilde{\overline{x^2}} - \tilde{\bar{x}}^2}.$$
7. $$\widetilde{S_0^2} = \frac{n\widetilde{\overline{y^2}} - n\tilde{\bar{y}}^2}{n-r}.$$

$$\widetilde{S^2} = \frac{n\widetilde{\overline{y^2}} - 2\tilde{\beta}_2 n\tilde{\bar{y}} - 2\tilde{\beta}_1 n\widetilde{\overline{xy}} + n\tilde{\beta}_2^2 + 2\tilde{\beta}_1\tilde{\beta}_2 n\tilde{\bar{x}} + \tilde{\beta}_1^2 n\widetilde{\overline{x^2}}}{n-r}.$$

$(\tilde{\theta}_0, \tilde{\theta}_1) = (\perp, \perp)$

if $\min(\widetilde{S_0^2}, (n\widetilde{\overline{x^2}} - n\tilde{\bar{x}}^2)/(n-1)) > 0$ **then**

  $\tilde{\theta}_0 = (\tilde{\bar{y}}, \tilde{\bar{x}}, \widetilde{\overline{x^2}}, \widetilde{S_0^2}, n)$

  $\tilde{\theta}_1 = (\tilde{\beta}_1, \tilde{\bar{x}}, \widetilde{\overline{x^2}}, \widetilde{S^2}, n)$

return $(\tilde{\theta}_0, \tilde{\theta}_1)$

---

slope of the line between the datapoints, we have

$$\mathbb{1}\left\{\frac{y_{i+1} - y_i}{x_{i+1} - x_i} > 0\right\} = \mathbb{1}\left\{\beta_1 + \frac{e_{i+1} - e_i}{x_{i+1} - x_i} > 0\right\} \sim \text{Bern}(1/2),$$

provided that $x_{i+1} \neq x_i$ (if $x_{i+1} = x_i$, then we set the result to a random Bern(1/2)). Note that this holds for any continuous distribution for the e_i's, not necessarily normal. There are simple DP tests to determine whether $p = 1/2$ given a dataset drawn from Bern(p) by computing a noisy sum of the values in the dataset and comparing it to a noisy threshold

Algorithm 3: ρ-zCDP procedure `DPBern`

Data: $X \in \mathbb{R}^{n \times 2}; Y \in \mathbb{R}^n$
Input: $n \in \mathbb{N}$; $\rho > 0$
Let $x_1, \ldots, x_n$ be the observed 1-D independent variables from X
Let $\tau : [n] \to [n]$ be a randomly chosen permutation
$s = 0$
$n_0 = \lfloor n/2 \rfloor$
for $i = 1 \ldots n_0$ **do**
  if $x_{\tau(n_0+i)} - x_{\tau(i)} \neq 0$ **then**
    $s = s + \mathbb{1}\left\{\frac{Y_{\tau(n_0+i)} - Y_{\tau(i)}}{x_{\tau(n_0+i)} - x_{\tau(i)}} > 0\right\}$
  else
    $r \sim \text{Bern}(1/2)$
    $s = s + r$
$s = s + \mathcal{N}(0, \frac{1}{2\rho})$
Let $N_{\alpha/2}$ and $N_{1-\alpha/2}$ denote the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\mathcal{N}(n_0/2, \frac{n_0}{4} + \frac{1}{2\rho})$ respectively
if $s \notin (N_{\alpha/2}, N_{1-\alpha/2})$ **then**
  return Reject the null
else
  return Fail to Reject the null

(determined by a normal approximation to a binomial distribution). The DP regression estimators of (Dwork and Lei, 2009; Alabi et al., 2022) also calculate the slopes between pairs of points, but then outputs a DP *median* of the results.

The resulting algorithm for privately testing a linear relationship is shown in Algorithm 3. We first group the points into $\lfloor n/2 \rfloor$ pairs. Then we calculate s, the number of slopes that are positive. We add noise to this estimate to satisfy ρ-zCDP and then use the noisy estimate to decide to reject the null. The noise to satisfy zCDP is $\mathcal{N}(0, \frac{1}{2\rho})$ whereas a normal approximation to $\text{Bin}(n_0, 1/2)$ is $\mathcal{N}(n_0/2, n_0/4)$. As a result, we can reject the null hypothesis iff the DP observed number of 1s is not in the $(\alpha/2, 1 - \alpha/2)$ quantiles of $\mathcal{N}(\frac{n_0}{2}, \frac{n_0}{4} + \frac{1}{2\rho})$ where α is the target significance level.

Lemma 6 *For any $\rho > 0$ and $n \in \mathbb{N}$, Algorithm 3 satisfies ρ-zCDP.*

Proof Algorithm 3 pairs the points into $n_0 = \lfloor n/2 \rfloor$ pairs. Note that this is a 1-Lipschitz transformation (i.e., changing one datapoint will change a single slope estimate) so that the differential privacy guarantees are preserved (by Definition 16 and Lemma 17 of (Alabi et al., 2022)).

A single datapoint (x_i, y_i) can affect the sum s by at most 1 so the global sensitivity is 1. By post-processing and by Proposition 1.6 in (Bun and Steinke, 2016), via the use of the Gaussian Mechanism, Algorithm 3 satisfies ρ-zCDP. ■

We note that a uniformly most powerful DP Bernoulli tester has been designed in (Awan and Slavkovic, 2020). Using this tester might yield better power than the tester described in Algorithm 3. However, that work is for pure DP (whereas we use zCDP) and the test is computationally slower.

4.3 Testing Mixture Models

As we will show experimentally, the best framework for the mixture model test depends on the properties of the dataset. This can be seen as conditional inference (Andrews et al., 2019). We now discuss our F-statistic and Kruskal-Wallis approaches.

4.3.1 F-Statistic

In the non-private case, we can use the following test statistic for testing mixtures in simple linear regression models:

$$\begin{aligned} T(X, Y, \hat{\beta}, \hat{\beta}^N, n, r, q) &= \left(\frac{n-r}{r-q}\right) \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2} \\ &= \frac{n_1 \overline{x^2}_1 n_2 \overline{x^2}_2}{S^2 n \overline{x^2}} (\hat{\beta}_1 - \hat{\beta}_2)^2. \end{aligned}$$

In Algorithm 4, we apply the Gaussian mechanism to calculate the DP sufficient statistics. $\widetilde{S_0^2}$ and $\widetilde{S^2}$ are DP estimates of the sampling error under the null and alternative hypothesis, respectively. In particular, $\widetilde{S_0^2}$ corresponds to an estimate of the sampling error when the groups have the same distributional properties.

Lemma 7 *For any $\rho, \Delta > 0$, Algorithm 4 satisfies ρ-zCDP.*

Proof This follows from Proposition 1.6 in (Bun and Steinke, 2016) via the use of the Gaussian Mechanism to ensure zCDP.

The composition and post-processing properties of zCDP (Lemmas 1.7 and 1.8 in (Bun and Steinke, 2016)) can then be applied. The computation of the following statistics is each done to satisfy $\rho/8$-zCDP: $\tilde{\bar{x}}_1, \tilde{\bar{x}}_2, \widetilde{\overline{x^2}}_1, \widetilde{\overline{x^2}}_2, \widetilde{\overline{xy}}_1, \widetilde{\overline{xy}}_2, \widetilde{\overline{y^2}}_1, \widetilde{\overline{y^2}}_2$. The other statistics computed are post-processed DP releases.

As a result, the entire procedure satisfies ρ-zCDP. ■

Instantiating Algorithm 1 for the F-statistic Mixture Tester: If the procedure `DPStats`$_M$ returns $(\perp, \perp)$, then we fail to reject the null. Otherwise, we use the returned statistics
$\tilde{\theta}_1 = (\tilde{\beta}_1, \tilde{\beta}_2, \widetilde{\overline{x^2}}_1, \widetilde{\overline{x^2}}_2, \widetilde{\overline{x^2}}, \widetilde{S^2}, n_1, n_2, n)$ to create the test statistic and use
$\tilde{\theta}_0 = (\tilde{\beta}_1, \tilde{\bar{x}}, \widetilde{\overline{x^2}}, \widetilde{S_0^2}, n_1, n_2, n)$ to simulate the null distributions. $P_{\tilde{\theta}_0}$ is instantiated as a normal distribution and used to generate ${}^k x_i$ distributed as $\mathcal{N}(\tilde{\bar{x}}, (n\widetilde{\overline{x^2}} - n\tilde{\bar{x}}_1^2)/(n-1))$ and to generate ${}^k y_i$ distributed as $\tilde{\beta}_1 {}^k x_i + e_i$, $e_i \sim \mathcal{N}(0, \widetilde{S_0^2})$ for either group 1 with size n_1 or group 2 with size $n - n_1$.

Algorithm 4: ρ-zCDP procedure $\texttt{DPStats}_M$

Data: $X \in \mathbb{R}^{n\times 2}, Y \in \mathbb{R}^n$

Input: integer $n_1, n \geq 2$; $r, q \in \mathbb{N}$; $\rho > 0, \Delta > 0$

Set $\rho_0 = \rho/8$ and $n_2 = n - n_1$. Then compute the following:

1. $\tilde{\bar{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i|_{-\Delta}^{\Delta} + \mathcal{N}(0, \frac{2\Delta^2}{\rho_0 n_1^2})$, $\tilde{\bar{x}}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n} x_i|_{-\Delta}^{\Delta} + \mathcal{N}(0, \frac{2\Delta^2}{\rho_0 n_2^2})$,
   $\tilde{\bar{x}} = n_1/n \cdot \tilde{\bar{x}}_1 + n_2/n \cdot \tilde{\bar{x}}_2$.

2. $\widetilde{\overline{x^2}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^2|_{0}^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho_0 n_1^2})$, $\widetilde{\overline{x^2}}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n} x_i^2|_{0}^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho_0 n_2^2})$,
   $\widetilde{\overline{x^2}} = n_1/n \cdot \widetilde{\overline{x^2}}_1 + n_2/n \cdot \widetilde{\overline{x^2}}_2$.

3. $\widetilde{\overline{xy}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i y_i|_{-\Delta^2}^{\Delta^2} + \mathcal{N}(0, \frac{2\Delta^4}{\rho_0 n_1^2})$, $\widetilde{\overline{xy}}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n} x_i y_i|_{-\Delta^2}^{\Delta^2} + \mathcal{N}(0, \frac{2\Delta^4}{\rho_0 n_2^2})$,
   $\widetilde{\overline{xy}} = n_1/n \cdot \widetilde{\overline{xy}}_1 + n_2/n \cdot \widetilde{\overline{xy}}_2$.

4. $\widetilde{\overline{y^2}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i^2|_{0}^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho_0 n_1^2})$, $\widetilde{\overline{y^2}}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n} y_i^2|_{0}^{\Delta^2} + \mathcal{N}(0, \frac{\Delta^4}{2\rho_0 n_2^2})$,
   $\widetilde{\overline{y^2}} = n_1/n \cdot \widetilde{\overline{y^2}}_1 + n_2/n \cdot \widetilde{\overline{y^2}}_2$.

5. $$\tilde{\beta}_1 = \frac{\widetilde{\overline{xy}}_1}{\widetilde{\overline{x^2}}_1}, \quad \tilde{\beta}_2 = \frac{\widetilde{\overline{xy}}_2}{\widetilde{\overline{x^2}}_2}, \quad \tilde{\beta} = n_1/n \cdot \tilde{\beta}_1 + n_2/n \cdot \tilde{\beta}_2.$$

6. $$\widetilde{S_0^2} = \frac{n\widetilde{\overline{y^2}} + n\tilde{\beta}^2 - 2n\widetilde{\overline{xy}}\tilde{\beta}}{n-r}.$$

$$\widetilde{S^2} = \frac{n_1\widetilde{\overline{y^2}}_1 + n_1\tilde{\beta}_1^2 - 2n_1\widetilde{\overline{xy}}_1\tilde{\beta}_1 + n_2\widetilde{\overline{y^2}}_2 + n_2\tilde{\beta}_2^2 - 2n_2\widetilde{\overline{xy}}_2\tilde{\beta}_2}{n-r}.$$

$(\tilde{\theta}_0, \tilde{\theta}_1) = (\perp, \perp)$

if $\min(\widetilde{S_0^2}, (n\widetilde{\overline{x^2}} - n\tilde{\bar{x}}^2)/(n-1)) > 0$ **then**

  $\tilde{\theta}_0 = (\tilde{\beta}_1, \tilde{\bar{x}}, \widetilde{\overline{x^2}}, \widetilde{S_0^2}, n_1, n_2, n)$

  $\tilde{\theta}_1 = (\tilde{\beta}_1, \tilde{\beta}_2, \widetilde{\overline{x^2}}_1, \widetilde{\overline{x^2}}_2, \widetilde{\overline{x^2}}, \widetilde{S^2}, n_1, n_2, n)$

return $(\tilde{\theta}_0, \tilde{\theta}_1)$

4.3.2 Nonparametric Tests via Kruskal-Wallis

Couch, Kazan, Shi, Bray, and Groce (Couch et al., 2019) present DP analogues of nonparametric hypothesis testing methods (which require little or no distributional assumptions). They find that the DP variant of the Kruskal-Wallis test statistic is more powerful than the DP version of the traditional parametric statistics for testing if two groups have the same medians. Here, we reduce our problem of testing mixture models to their problem of testing if groups share the same median. The reduction is as follows: Given two datasets $(\mathbf{x}^1, \mathbf{y}^1)$ and $(\mathbf{x}^2, \mathbf{y}^2)$, each of size n_1 and n_2 respectively, we wish to test if the slopes are equal. We randomly match all pairs of points in $(\mathbf{x}^1, \mathbf{y}^1)$ and obtain at most $n_1/2$ slopes in s_1. We do the same for the second group $(\mathbf{x}^2, \mathbf{y}^2)$ to obtain $n_2/2$ slopes in s_2. Then we compute the mean of ranks of elements in s_1 and s_2 as $\bar{r}_1$ and $\bar{r}_2$ respectively. Next, we compute the Kruskal-Wallis absolute value test statistic h from (Couch et al., 2019) and release a perturbed version satisfying zCDP. We can use the Monte Carlo testing framework in Algorithm 1 and use Algorithm 5 to compute the test statistics. Under the null, the slopes in s_1 and s_2 would have similar ranks so we choose uniform random numbers in some interval. We then decide to reject or fail to reject the null, based on the distribution of test statistics obtained via this process and its relation to the statistic computed on the observed data.

Lemma 8 *For any $\rho > 0$ and even n, Algorithm 5 satisfies ρ-zCDP.*

Proof Algorithm 5 randomly pairs all n_1 pairs of points in group 1 (to obtain slopes s_1 of size $n_1/2$) and pairs all n_2 pairs in group 2 (to obtain slopes s_2 of size $n_2/2$). Note that this is a 1-Lipschitz transformation so that the differential privacy guarantees are preserved (by Definition 16 and Lemma 17 of (Alabi et al., 2022)).

Then we proceed to use the DP Kruskal-Wallis absolute value test statistic with sensitivity of 8 (as shown in Theorem 3.4 of (Couch et al., 2019)). By Proposition 1.6 in (Bun and Steinke, 2016), via the use of the Gaussian Mechanism, the procedure satisfies ρ-zCDP. ■

Instantiating Algorithm 1 for the Kruskal-Wallis Mixture Tester: We use the returned statistic $\tilde{\theta}_1 = (h)$ as the sole statistic. In this case, T is the identity function. $\tilde{\theta}_0$ is taken to be null. $P_{\tilde{\theta}_0}$ generates ${}^k x_i$ and ${}^k y_i$ (for the 2 groups) independently and uniformly at random in a fixed interval (say $[-5, 5]$). Although this distribution may be very different from the actual data distribution, the distribution of ranks of the slopes will be identical to that under the null, ensuring that $T(({}^k x, {}^k y))$ has the right distribution.

5. Differentially Private F-statistic

In this section, we will show that the DP F-statistic converges, in distribution, to the asymptotic distribution of the F-statistic. The focus will be on showing results for Algorithm 2 but a similar route can be used to obtain analogous results for Algorithm 4. Recall that Algorithm 2 is an instantiation of the DP F-statistic for testing a linear relationship while Algorithm 4 is for testing mixtures.

T_n is the non-private F-statistic while $\tilde{T}_n$ is the DP F-statistic constructed from DP sufficient statistics obtained via Algorithm 2. The main theorem in this section is Theorem 9, which shows the convergence, in distribution, of $\tilde{T}_n$ to the asymptotic distribution of T_n, the chi-squared distribution. As a corollary, the statistical power of $\tilde{T}_n$ converges to the statistical power of T_n. While Theorem 9 is specialized to the simple linear regression setting (i.e., $p = 2$), it can easily be extended to multiple linear regression.

Theorem 9 *Let $\sigma_e > 0$, $r = p = 2, q = 1$, and $\beta \in \mathbb{R}^p$. For every $n \in \mathbb{N}$ with $n > r$, let $X_n \in \mathbb{R}^{n \times p}$ be the design matrix where the first column is an all-ones vector and the second*

Algorithm 5: ρ-zCDP procedure `DPKW`

Data: $X \in \mathbb{R}^{n\times 2}; Y \in \mathbb{R}^n$
Input: Even $n_1, n \in \mathbb{N}$; $\rho > 0$
Let $x_1, \ldots, x_n$ be the observed 1-D independent variables from X
Let $\tau : [n] \to [n]$ be a randomly chosen permutation
$s_1 = \{\}$
for $i = 1 \ldots n_1/2$ **do**
  $s_1 = s_1 \bigcup \{\frac{Y_{\tau(n_1/2+i)} - Y_{\tau(i)}}{x_{\tau(n_1/2+i)} - x_{\tau(i)}}\}$
$n_2 = n - n_1$
$s_2 = \{\}$
$e = n_1 + n_2/2$
for $i = 1 \ldots n_2/2$ **do**
  $s_2 = s_2 \bigcup \{\frac{Y_{\tau(e+i)} - Y_{\tau(i+n_1)}}{x_{\tau(e+i)} - x_{\tau(i+n_1)}}\}$
Let $r : \mathbb{R}^m \to [m]$ be rank-computing function on any m elements
Compute s by appending (in an order-preserving manner) s_2 to s_1
Compute $\bar{r}_1$, mean of ranks of s_1 in $r(s)$
Compute $\bar{r}_2$, mean of ranks of s_2 in $r(s)$
Compute $h = \frac{4(n-1)}{n^2}\left(n_1|\bar{r}_1 - \frac{n+1}{2}| + n_2|\bar{r}_2 - \frac{n+1}{2}|\right)$
return null, $h + \mathcal{N}(0, 8^2/(2\rho))$

column is $(x_1, \ldots, x_n)^T$. Let $\Delta = \Delta_n > 0$ be a sequence of clipping bounds, $\rho = \rho_n > 0$ be a sequence of privacy parameters, and $\eta_n^2 = \frac{\|X_n\beta - X_n\beta^N\|^2}{\sigma_e^2}$. Under the multivariate linear regression model, $Y_n \sim \mathcal{N}(X_n\beta, \sigma_e^2 I_{n\times n})$. Let $\tilde{\beta}$ and $\tilde{\beta}^N$ be the DP least-squares estimate of β, obtained in Algorithm 2, under the alternative and null hypotheses, respectively. Let $\tilde{T} = \tilde{T}_n$ be the DP F-statistic computed from DP sufficient statistics via Algorithm 2 and Equation (4). Suppose the following conditions hold:

1. $\exists c_x, c_{x^2} \in \mathbb{R}$ *such that* $\bar{x} \to c_x$, $\overline{x^2} \to c_{x^2}$, *and* $c_{x^2} > c_x^2$,
2. $\exists \eta \in \mathbb{R}$ *such that* $\eta_n^2 \to \eta^2$,
3. $\frac{\Delta_n^2}{\rho_n n}, \frac{\Delta_n^4}{\rho_n n} \to 0$,
4. $\mathbb{P}[\exists i \in [n], y_i \notin [-\Delta_n, \Delta_n]] \to 0$ *and* $\forall i \in [n], x_i \in [-\Delta_n, \Delta_n]$.

Then we obtain the following results:

1. *Under the null hypothesis:* $\tilde{\beta}^N = \tilde{\beta}_n^N \xrightarrow{P} \beta$,
2. *Under the alternative hypothesis:* $\tilde{\beta} = \tilde{\beta}_n \xrightarrow{P} \beta$,
3. $\tilde{T} = \tilde{T}_n \xrightarrow{D} \frac{\chi^2_{r-q}(\eta^2)}{r-q}$.

The condition that $\mathbb{P}[\exists i \in [n], y_i \notin [-\Delta_n, \Delta_n]] \to 0$ (Condition 4 in Theorem 9), holds by a Gaussian tail bound (Claim 5.1), if $\exists k > 0$ such that for all $i \in [n]$, $\Delta_n \geq |\beta_1 x_i + \beta_2| + \sigma_e \sqrt{\log 2n^k}$.

First, we will prove convergence results for sufficient statistics used to construct the non-private F-statistic T_n, in our setting. Then we will show convergence results for DP sufficient statistics used to construct the private F-statistic $\tilde{T}_n$. Finally, we will combine these previous results to show Theorem 9.

5.1 Convergence of Non-private Sufficient Statistics

In Equation (4), the non-private F-statistic is given as

$$T = T_n = \frac{n-r}{r-q} \cdot \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2} = \frac{n-r}{r-q} \cdot \frac{\|X_n\hat{\beta} - X_n\hat{\beta}^N\|^2}{\|Y_n - X_n\hat{\beta}\|^2}.$$

We start by writing this F-statistic, in an equivalent form, in terms of quantities that we will show are convergent:

Lemma 10 *Suppose that $\sigma_e > 0$, $p \in \mathbb{N}$, and $\beta \in \mathbb{R}^p$. Let $X = X_n \in \mathbb{R}^{n \times p}$ be the full-rank design matrix (as in Equation (1)) and $Y = Y_n \sim \mathcal{N}(X_n\beta, \sigma_e^2 I_{n\times n})$. Also, let $\hat{\beta}$ and $\hat{\beta}^N$ be the non-private least-squares estimate of β under the alternative and null hypotheses, respectively.*

Define the following quantities:

$$\hat{E}_n = \left(\frac{X_n^T X_n}{n}\right)^{1/2} \in \mathbb{R}^{p\times p}, \quad \hat{F}_n = \frac{X_n^T Y_n}{n} \in \mathbb{R}^p, \quad \hat{G}_n = \frac{Y_n^T Y_n}{n} \in \mathbb{R}.$$

Then the test statistic T_n from Equation (4) can be re-written as

$$T_n = \frac{n-r}{r-q} \cdot \frac{\|\sqrt{n}\hat{E}_n(\hat{\beta} - \hat{\beta}^N)\|^2}{n(\hat{\beta}^T \hat{E}_n^2 \hat{\beta} - 2\hat{\beta}^T \hat{F}_n + \hat{G}_n)}, \tag{7}$$

for any $n, r, q \in \mathbb{N}$ such that $q < r$.

Proof [Proof of Lemma 10] First, note that $\hat{E}_n \in \mathbb{R}^{p\times p}$ (i) exists because $X_n^T X_n$ is positive definite, (ii) is unique since its square is positive definite (Horn and Johnson, 2012).

$$\begin{aligned}\|X_n\hat{\beta} - X_n\hat{\beta}^N\|^2 &= (\hat{\beta} - \hat{\beta}^N)^T X_n^T X_n (\hat{\beta} - \hat{\beta}^N)\\ &= \sqrt{n}(\hat{\beta} - \hat{\beta}^N)^T \hat{E}_n^T \sqrt{n}\hat{E}_n(\hat{\beta} - \hat{\beta}^N)\\ &= \|\sqrt{n}\hat{E}_n(\hat{\beta} - \hat{\beta}^N)\|^2.\end{aligned}$$

Next,

$$\begin{aligned}\|Y_n - X_n\hat{\beta}\|^2 &= (Y_n - X_n\hat{\beta})^T (Y_n - X_n\hat{\beta})\\ &= Y_n^T Y_n - Y_n^T X_n\hat{\beta} - \hat{\beta}^T X_n^T Y_n + \hat{\beta}^T X_n^T X_n \hat{\beta}\\ &= Y_n^T Y_n + \hat{\beta}^T X_n^T X_n \hat{\beta} - 2\hat{\beta}^T X_n^T Y_n\\ &= n(\hat{\beta}^T \hat{E}_n^2 \hat{\beta} - 2\hat{\beta}^T \hat{F}_n + \hat{G}_n).\end{aligned}$$

■

It will be easier to use Equation (7) as an equivalent form of the F-statistic to prove convergence results. An analogous representation will be used to prove convergence results for the DP F-statistic.

In the case of testing a linear relationship (as in Section 3.1) in simple linear regression (i.e., where $p = 2$ and the columns of X are the all-ones vector and $(x_1, \ldots, x_n)^T$),

$$X^T X = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\overline{x^2} \end{pmatrix}, \quad X^T Y = \begin{pmatrix} n\bar{y} \\ n\overline{xy} \end{pmatrix}, \quad Y^T Y = \sum_{i=1}^{n} y_i^2,$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_1 \end{pmatrix}, \quad \hat{\beta}^N = \begin{pmatrix} \bar{y} \\ 0 \end{pmatrix},$$

$$\widehat{\sigma_x^2} = \overline{x^2} - \bar{x}^2.$$

In this case, it can be verified that $\hat{E}_n, \hat{F}_n, \hat{G}_n$ is:

$$\begin{aligned} \hat{E}_n &= \frac{1}{\sqrt{\overline{x^2} + 1 + 2\sqrt{\overline{x^2} - \bar{x}^2}}} \begin{pmatrix} 1 + \sqrt{\overline{x^2} - \bar{x}^2} & \bar{x} \\ \bar{x} & \overline{x^2} + \sqrt{\overline{x^2} - \bar{x}^2} \end{pmatrix}, \\ &= \frac{1}{\sqrt{\overline{x^2} + 1 + 2\sqrt{\widehat{\sigma_x^2}}}} \begin{pmatrix} 1 + \sqrt{\widehat{\sigma_x^2}} & \bar{x} \\ \bar{x} & \overline{x^2} + \sqrt{\widehat{\sigma_x^2}} \end{pmatrix}, \\ \hat{F}_n &= \frac{X^T Y}{n} = \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}, \quad \hat{G}_n = \frac{Y^T Y}{n} \overset{\text{def}}{=} \overline{y^2}. \end{aligned} \tag{8}$$

Next, we proceed to show non-private convergence results that will be pivotal to our final result. We will crucially rely on the Gaussian tail bound, the normality of $\hat{\beta}, \hat{\beta}^N$, and Corollary 12.

Lemma 11 *For every sequence of clipping bounds $\Delta = \Delta_n > 0$ and sequence of privacy parameters $\rho = \rho_n > 0$, under the conditions of Theorem 9:*

(1) $\exists c_y \in \mathbb{R}$ such that $\bar{y} \xrightarrow{P} c_y$,

(2) $\exists c_a, c_{xy} \in \mathbb{R}$ such that $\overline{xy} \xrightarrow{P} c_{xy}$, $\overline{xy} - \bar{x} \cdot \bar{y} \xrightarrow{P} c_a$,

(3) $\exists c_b \neq 0$ such that $\widehat{\sigma_x^2} \xrightarrow{P} c_b$,

(4) $\exists$ unique positive-definite $C^{1/2} \in \mathbb{R}^{2\times 2}$ such that $\hat{E}_n \to C^{1/2}$,

(5) $\hat{F}_n \xrightarrow{P} (c_y \quad c_{xy})^T$,

(6) $\exists c_{y^2} \in \mathbb{R}$ such that $\hat{G}_n = \frac{Y^T Y}{n} \xrightarrow{P} c_{y^2}$,

(7) Normality of $\hat{\beta}$: $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma_e^2 (X_n^T X_n)^{-1}\right)$; Consistency of $\hat{\beta}$: $\hat{\beta} \xrightarrow{P} \beta$.

A self-contained proof of Lemma 11 can be found in (Alabi and Vadhan, 2022).

To prove our main theorem, we will make use of the following tools: the Gaussian tail bound and Slutsky's Theorem, which we state below.

Claim 5.1 (Gaussian Tail Bound) *Let Z be a standard normal random variable with mean 0 and variance 1. i.e., $Z \sim \mathcal{N}(0,1)$. Then*

$$\mathbb{P}[|Z| > t] \leq 2\exp(-t^2/2),$$

for every $t > 0$.

By the Gaussian tail bound, any Gaussian random variable (such as the DP estimates) converges, in probability, to the asymptotic distributions of the estimates without Gaussian noise added as long as the variance goes to 0 (Corollary 12):

Corollary 12 *Let $N_n \sim \mathcal{N}(0, \sigma_n^2)$ where $\sigma_n \to 0$, then $N_n \xrightarrow{P} 0$.*

Corollary 12 follows from the definition of convergence in probability and the Gaussian tail bound (Claim 5.1).

Next, we introduce Slutsky's Theorem which will be crucial to combining individual convergence results to show more general results:

Theorem 13 (Slutsky's Theorem, see (Gut, 2013)) *Let $\{W_n\}, \{Z_n\}$ be a sequence of random vectors and W be a random vector. If $W_n \xrightarrow{D} W$ and $Z_n \xrightarrow{P} c$ for a constant $c \in \mathbb{R}$, then as $n \to \infty$:*

1. $W_n \cdot Z_n \xrightarrow{D} Wc$,
2. $W_n + Z_n \xrightarrow{D} W + c$,
3. $W_n / Z_n \xrightarrow{D} W/c$ *as long as* $c \neq 0$.

Now, we will show that the DP statistics converge, either in probability or distribution, to the distributions of their corresponding non-DP statistics.

5.2 Convergence of Differentially Private Sufficient Statistics

The DP F-statistic is constructed via Algorithm 2 and Equation (4). We start by rewriting the DP F-statistic analogously to Lemma 10:

Lemma 14 *Suppose that $\sigma_e > 0$, $p \in \mathbb{N}$, and $\beta \in \mathbb{R}^p$. Let $X = X_n \in \mathbb{R}^{n \times p}$ be the full-rank design matrix (as in Equation (1)) and $Y = Y_n \sim \mathcal{N}(X_n\beta, \sigma_e^2 I_{n\times n})$. Also, let $\tilde{\bar{x}}, \tilde{\bar{y}}$, $\widetilde{\overline{x^2}}, \widetilde{\overline{xy}}$, $\widetilde{\overline{y^2}}$, $\tilde{\beta}_1$, and $\tilde{\beta}_2$ be as computed in Algorithm 2.*

Define the following quantities:

$$\widetilde{\sigma_x^2} \stackrel{\text{def}}{=} \widetilde{x^2} - \tilde{\bar{x}}^2, \tag{9}$$

$$\tilde{E}_n = \left(\frac{\widetilde{X^T X}}{n}\right)^{1/2} = \frac{1}{\sqrt{\widetilde{x^2} + 1 + 2\sqrt{\widetilde{\sigma_x^2}}}} \begin{pmatrix} 1 + \sqrt{\widetilde{\sigma_x^2}} & \tilde{\bar{x}} \\ \tilde{\bar{x}} & \widetilde{x^2} + \sqrt{\widetilde{\sigma_x^2}} \end{pmatrix},$$

$$\tilde{F}_n = \frac{\widetilde{X^T Y}}{n} = \begin{pmatrix} \tilde{\bar{y}} \\ \widetilde{xy} \end{pmatrix}, \quad \tilde{G}_n = \widetilde{y^2}, \tag{10}$$

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_2 \\ \tilde{\beta}_1 \end{pmatrix}, \quad \tilde{\beta}^N = \begin{pmatrix} \tilde{\bar{y}} \\ 0 \end{pmatrix}. \tag{11}$$

where we take $\sqrt{\widetilde{\sigma_x^2}}$ to be the square root of $\widetilde{\sigma_x^2}$ with non-negative real and imaginary parts.

Furthermore, if $\tilde{T}_n = T(\tilde{\theta}_1)$ is the test statistic obtained via statistics computed in Algorithm 2 and via Equation (4), then $\tilde{T}_n$ can be re-written as

$$\tilde{T} = \tilde{T}_n = \frac{n-r}{r-q} \cdot \frac{\|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2}{n(\tilde{\beta}^T \tilde{E}_n^2 \tilde{\beta} - 2\tilde{\beta}^T \tilde{F}_n + \tilde{G}_n)}, \tag{12}$$

for any $n, r, q \in \mathbb{N}$ such that $q < r$.

Proof [Proof of Lemma 14]

$$\begin{aligned}(\tilde{\beta} - \tilde{\beta}^N)^T \widetilde{X_n^T X_n} (\tilde{\beta} - \tilde{\beta}^N) &= \sqrt{n}(\tilde{\beta} - \tilde{\beta}^N)^T \tilde{E}_n^T \sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N) \\ &= \|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2.\end{aligned}$$

Next,

$$\begin{aligned}\widetilde{Y_n^T Y_n} - 2\tilde{\beta}^T \widetilde{X_n^T Y_n} + \tilde{\beta}^T \widetilde{X_n^T X_n} \tilde{\beta} &= n\tilde{G}_n - 2n\tilde{\beta}^T \tilde{F}_n + n\tilde{\beta}^T \hat{E}_n^2 \tilde{\beta} \\ &= n(\tilde{\beta}^T \tilde{E}_n^2 \tilde{\beta} - 2\tilde{\beta}^T \tilde{F}_n + \tilde{G}_n).\end{aligned}$$

■

We now introduce two helper lemmas that are useful for showing later results. The first uses a hybrid-type argument to show a $1/f(n)$ rate of convergence of a ratio of random variables. The second can be used to show that if the difference of two random variables converge to 0, then as long as they converge to a non-zero constant, the difference of their square root converge to 0.

Lemma 15 *Let $A_n, B_n, \widetilde{A}_n, \widetilde{B}_n$ be random variables such that:*

1. *For constants $c_1, c_2 \in \mathbb{R}$, $c_2 \neq 0$: $A_n \xrightarrow{P} c_1$, $B_n \xrightarrow{P} c_2$,*
2. *For function $f(n)$: $f(n)(\widetilde{A}_n - A_n) \xrightarrow{P} 0$ and $f(n)(\widetilde{B}_n - B_n) \xrightarrow{P} 0$.*

Then:

$$f(n)\left(\frac{\widetilde{A}_n}{\widetilde{B}_n}-\frac{A_n}{B_n}\right)\xrightarrow{P}0.$$

Proof [Proof of Lemma 15] We use a hybrid-type argument. We write

$$f(n)\left(\frac{\widetilde{A}_n}{\widetilde{B}_n}-\frac{A_n}{B_n}\right)=f(n)\left(\frac{\widetilde{A}_n}{\widetilde{B}_n}-\frac{\widetilde{A}_n}{B_n}+\frac{\widetilde{A}_n}{B_n}-\frac{A_n}{B_n}\right).$$

Then,

$$\begin{aligned}f(n)\left(\frac{\widetilde{A}_n}{\widetilde{B}_n}-\frac{\widetilde{A}_n}{B_n}\right)&=f(n)\left(\frac{\widetilde{A}_nB_n-\widetilde{A}_n\widetilde{B}_n}{\widetilde{B}_nB_n}\right)\\&=\widetilde{A}_nf(n)\left(\frac{B_n-\widetilde{B}_n}{\widetilde{B}_nB_n}\right)\\&\xrightarrow{P}0,\end{aligned}$$

since $\widetilde{A}_n\xrightarrow{P}c_1$, $f(n)(B_n-\widetilde{B}_n)\xrightarrow{P}0$, and by Slutsky's Theorem $B_n\widetilde{B}_n\xrightarrow{P}c_2^2\neq 0$.

Also,

$$f(n)\left(\frac{\widetilde{A}_n}{B_n}-\frac{A_n}{B_n}\right)=f(n)\left(\frac{\widetilde{A}_n-A_n}{B_n}\right)\xrightarrow{P}0,$$

since $f(n)(\widetilde{A}_n-A_n)\xrightarrow{P}0$, $B_n\xrightarrow{P}c_2\neq 0$ so that the result follows by a routine application of Slutsky's Theorem.

As a result, $f(n)\left(\frac{\widetilde{A}_n}{\widetilde{B}_n}-\frac{A_n}{B_n}\right)\xrightarrow{P}0$.

■

Lemma 16 *Let $A_n,\widetilde{A}_n$ be random variables such that:*

1. *For constant $c\in\mathbb{R}$, $c\neq 0$: $\quad A_n\xrightarrow{P}c$,*
2. *For function $f(n)$: $\quad f(n)(\widetilde{A}_n-A_n)\xrightarrow{P}0$.*

Then:

$$f(n)(\widetilde{A}_n^{1/2}-A_n^{1/2})\xrightarrow{P}0.$$

Proof [Proof of Lemma 16] Throughout, we take square roots in which both the real and imaginary parts are non-negative.

Recall that by difference of two squares:

$$a^{1/2}-b^{1/2}=\frac{a-b}{a^{1/2}+b^{1/2}},$$

for any $a, b \in \mathbb{C}$.

Then by Slutsky's Theorem:

$$f(n)(\widetilde{A}_n^{1/2} - A_n^{1/2}) = \frac{f(n)(\widetilde{A}_n - A_n)}{\widetilde{A}_n^{1/2} + A_n^{1/2}} \xrightarrow{P} 0,$$

where $\widetilde{A}_n^{1/2}, A_n^{1/2} \xrightarrow{P} c^{1/2}$.

■

We will show that the DP regression coefficients converge to the true coefficients. i.e., $\tilde{\beta} \xrightarrow{P} \beta$. But we begin with showing convergence of the constituent DP sufficient statistics.

Lemma 17 *For every sequence of clipping bounds $\Delta = \Delta_n > 0$ and sequence of privacy parameters $\rho = \rho_n > 0$, in Algorithm 2, under the conditions of Theorem 9:*

(1) $\sqrt{n}|\tilde{\bar{x}} - \bar{x}| \xrightarrow{P} 0$, $\tilde{\bar{x}} \xrightarrow{P} c_x$,

(2) $\sqrt{n}|\tilde{\bar{y}} - \bar{y}| \xrightarrow{P} 0$, $\tilde{\bar{y}} \xrightarrow{P} c_y$,

(3) $\sqrt{n}|\widetilde{\overline{x^2}} - \overline{x^2}| \xrightarrow{P} 0$, $\widetilde{\overline{x^2}} \xrightarrow{P} c_{x^2}$,

(4) $\sqrt{n}|\widetilde{\overline{xy}} - \overline{xy}| \xrightarrow{P} 0$, $\widetilde{\overline{xy}} \xrightarrow{P} c_{xy}$,

(5) $\sqrt{n}|\tilde{\bar{x}}^2 - \bar{x}^2| \xrightarrow{P} 0$,

(6) $\sqrt{n}|\tilde{\bar{x}}\tilde{\bar{y}} - \bar{x} \cdot \bar{y}| \xrightarrow{P} 0$,

(7) $\exists c_a \in \mathbb{R}$, $\widetilde{\overline{xy}} - \tilde{\bar{x}}\tilde{\bar{y}} \xrightarrow{P} c_a$,

(8) $\exists c_b \neq 0$, $\widetilde{\sigma_x^2} \xrightarrow{P} c_b$,

(9) $\sqrt{n}(\widetilde{\sigma_x^2} - \widehat{\sigma_x^2}) \xrightarrow{P} 0$,

(10) $\exists C^{1/2} \in \mathbb{R}^{2\times 2}$ *such that* $\sqrt{n}(\tilde{E}_n - \hat{E}_n) \xrightarrow{P} 0$, $\tilde{E}_n \xrightarrow{P} C^{1/2}$,

(11) $\tilde{F}_n \xrightarrow{P} (c_y \quad c_{xy})^T$,

(12) $\exists c_{y^2} \in \mathbb{R}$ *such that* $\tilde{G}_n \xrightarrow{P} c_{y^2}$,

where the constant scalars and matrix $c_x, c_y, c_{x^2}, c_{xy}, c_{y^2}, c_a, c_b, C^{1/2}$ are the same as the ones defined in Lemma 11.

Proof [Proof of Lemma 17] Define

1. $\breve{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i|_{-\Delta_n}^{\Delta_n}$,

2. $\breve{y} = \frac{1}{n}\sum_{i=1}^{n} y_i|_{-\Delta_n}^{\Delta_n}$,

3. $\overbrace{\overline{x^2}} = \frac{1}{n}\sum_{i=1}^{n} x_i^2|_{0}^{\Delta_n^2}$,

4. $\overbrace{\overline{xy}} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i|_{-\Delta_n^2}^{\Delta_n^2}$, and

5. $\overbrace{\overline{y^2}} = \sum_{i=1}^{n} y_i^2|_{0}^{\Delta_n^2}$.

Then $\tilde{\breve{x}} = \breve{x} + N_1$, $\tilde{\breve{y}} = \breve{y} + N_2$, $\widetilde{\overline{x^2}} = \overbrace{\overline{x^2}} + N_3$, $\widetilde{\overline{xy}} = \overbrace{\overline{xy}} + N_4$, $\widetilde{\overline{y^2}} = \overbrace{\overline{y^2}} + N_5$ where $N_1, N_2 \sim \mathcal{N}(0, \frac{2\Delta^2}{\rho n^2})$, $N_3 \sim \mathcal{N}(0, \frac{\Delta^4}{2\rho n^2})$, $N_4 \sim \mathcal{N}(0, \frac{2\Delta^4}{\rho n^2})$, and $N_5 \sim \mathcal{N}(0, \frac{\Delta^4}{2\rho})$.

By conditions of Theorem 9, $\frac{\Delta_n^2}{\rho_n n} \to 0$, $\frac{\Delta_n^4}{\rho_n n} \to 0$ so that by Corollary 12, $\sqrt{n}|N_1| \xrightarrow{P} 0$, $\sqrt{n}|N_2| \xrightarrow{P} 0$, $\sqrt{n}|N_3| \xrightarrow{P} 0$, $\sqrt{n}|N_4| \xrightarrow{P} 0$, and $\frac{\sqrt{n}}{n}|N_5| \xrightarrow{P} 0$ since $\sqrt{n}\mathcal{N}(0, \frac{2\Delta^2}{\rho n^2}) \sim \mathcal{N}(0, \frac{2\Delta^2}{\rho n})$, $\sqrt{n}\mathcal{N}(0, \frac{\Delta^4}{2\rho n^2}) \sim \mathcal{N}(0, \frac{\Delta^4}{2\rho n})$, and $\frac{\sqrt{n}}{n}\mathcal{N}(0, \frac{2\Delta^4}{\rho}) \sim \mathcal{N}(0, \frac{2\Delta^4}{\rho n})$.

(1): By assumption, for all $i \in [n]$, $x_i \in [-\Delta_n, \Delta_n]$. Thus, $\bar{x} = \breve{x}$ so that $\tilde{\breve{x}} = \bar{x} + N_1$. Then,

$$\sqrt{n}|\tilde{\breve{x}} - \bar{x}| \le \sqrt{n}|N_1| + \sqrt{n}|\breve{x} - \bar{x}| \xrightarrow{P} 0$$

by Slutsky's Theorem. As a corollary, $\tilde{\breve{x}} \xrightarrow{P} c_x$ by Lemma 25 and the assumption in Theorem 9 that $\bar{x} \to c_x$.

(2): The proof that $\sqrt{n}|\bar{y} - \tilde{\breve{y}}| \xrightarrow{P} 0$ is very similar: observe that by the assumptions of Theorem 9:

$$\begin{aligned} \mathbb{P}[|\bar{y} - \breve{y}| > 0] &\le \mathbb{P}[\exists i \in [n], y_i \notin [-\Delta_n, \Delta_n]] \qquad (13)\\ &\to 0. \qquad (14)\end{aligned}$$

Thus, $\sqrt{n}|\breve{y} - \bar{y}| \xrightarrow{P} 0$. Combining with $\sqrt{n}|N_2| \xrightarrow{P} 0$, by the triangle inequality, $\sqrt{n}|\bar{y} - \tilde{\breve{y}}| \xrightarrow{P} 0$. As a corollary, $\tilde{\breve{y}} \xrightarrow{P} c_y$ by Lemma 25 and the assumption in Theorem 9 that $\bar{x} \to c_x$.

(3): To show $\sqrt{n}|\widetilde{\overline{x^2}} - \overline{x^2}| \xrightarrow{P} 0$, we proceed in an analogous way: using the assumption that $\mathbb{P}[\exists i \in [n], x_i \notin [-\Delta_n, \Delta_n]] = 0$, we obtain that $\mathbb{P}[\exists i \in [n], x_i^2 \notin [0, \Delta_n^2]] = 0$ so that $\sqrt{n}|\overline{x^2} - \overbrace{\overline{x^2}}| \xrightarrow{P} 0$. Combining with $\sqrt{n}|N_3| \xrightarrow{P} 0$, by the triangle inequality, $\sqrt{n}|\widetilde{\overline{x^2}} - \overline{x^2}| \xrightarrow{P} 0$. As a corollary, $\widetilde{\overline{x^2}} \xrightarrow{P} c_{x^2}$ by Lemma 25 and the assumption in Theorem 9 that $\overline{x^2} \to c_{x^2}$.

(4): In a similar fashion, $\sqrt{n}|\widetilde{\overline{xy}} - \overline{xy}| \xrightarrow{P} 0$: using the assumptions

$$\mathbb{P}[\exists i \in [n], x_i \notin [-\Delta_n, \Delta_n]] = 0, \quad \mathbb{P}[\exists i \in [n], y_i \notin [-\Delta_n, \Delta_n]] \to 0,$$

we have that

$$\begin{aligned} &\mathbb{P}[\exists i \in [n], x_i y_i \notin [-\Delta_n^2, \Delta_n^2]]\\ &\le \mathbb{P}[\exists i \in [n], x_i \notin [-\Delta_n, \Delta_n]] + \mathbb{P}[\exists i \in [n], y_i \notin [-\Delta_n, \Delta_n]]\\ &\to 0,\end{aligned}$$

so that $\sqrt{n}|\overline{xy} - \overset{\smile}{\overline{xy}}| \xrightarrow{P} 0$. Combining with $\sqrt{n}|N_4| \xrightarrow{P} 0$, by the triangle inequality, $\sqrt{n}|\widetilde{\overline{xy}} - \overline{xy}| \xrightarrow{P} 0$. By Lemma 11, $\overline{xy} \xrightarrow{P} c_{xy}$. Then by Lemma 25, $\widetilde{\overline{xy}} \xrightarrow{P} c_{xy}$.

(5): Next we show $\sqrt{n}|\tilde{\bar{x}}^2 - \bar{x}^2| \xrightarrow{P} 0$: $\sqrt{n}(\tilde{\bar{x}}^2 - \bar{x}^2) = \sqrt{n}(\tilde{\bar{x}} - \bar{x})(\tilde{\bar{x}} + \bar{x})$. Since, $\tilde{\bar{x}}, \bar{x} \xrightarrow{P} c_x$, we have $(\tilde{\bar{x}} + \bar{x}) \xrightarrow{P} 2c_x$, $(\tilde{\bar{x}} - \bar{x}) \xrightarrow{P} 0$ so that by Slutsky's Theorem, $\sqrt{n}|\tilde{\bar{x}}^2 - \bar{x}^2| \xrightarrow{P} 0$

(6): In a similar fashion, $\sqrt{n}|\tilde{\bar{x}}\tilde{\bar{y}} - \bar{x} \cdot \bar{y}| \xrightarrow{P} 0$: $\tilde{\bar{x}} = \breve{\bar{x}} + N_1 = \bar{x} + N_1$, since $\forall i \in [n], x_i \in [-\Delta_n, \Delta_n]$. Then,

$$\begin{aligned}\sqrt{n}(\tilde{\bar{x}}\tilde{\bar{y}} - \bar{x} \cdot \bar{y}) &= \sqrt{n}\left[(\bar{x} + N_1)\tilde{\bar{y}} - \bar{x} \cdot \bar{y}\right] \\ &= \sqrt{n}N_1\tilde{\bar{y}} + \bar{x}\sqrt{n}(\tilde{\bar{y}} - \bar{y}) \\ &\xrightarrow{P} 0,\end{aligned}$$

since $\sqrt{n}(\tilde{\bar{y}} - \bar{y}) \xrightarrow{P} 0$, $\tilde{\bar{y}} \xrightarrow{P} c_y$, $\sqrt{n}N_1 \xrightarrow{P} 0$.

(7): Next, we show that $\widetilde{\overline{xy}} - \tilde{\bar{x}}\tilde{\bar{y}} \xrightarrow{P} c_a \in \mathbb{R}$: Follows by Slutsky's Theorem since $\widetilde{\overline{xy}} \xrightarrow{P} c_{xy}$, $\tilde{\bar{x}} \xrightarrow{P} c_x$, and $\tilde{\bar{y}} \xrightarrow{P} c_y$.

(8): In a similar fashion, $\widetilde{\sigma_x^2} \xrightarrow{P} c_b$: This follows by Slutsky's Theorem since $\widetilde{\overline{x^2}} \xrightarrow{P} c_{x^2}$, $\tilde{\bar{x}} \xrightarrow{P} c_x$.

(9): $\sqrt{n}(\widetilde{\sigma_x^2} - \widehat{\sigma_x^2}) \xrightarrow{P} 0$ follows from parts (3) and (5).

(10): By Lemma 16,

$$\sqrt{n}\left(\sqrt{\widetilde{\sigma_x^2}} - \sqrt{\widehat{\sigma_x^2}}\right) \xrightarrow{P} 0,$$

since $\sqrt{n}\left(\widetilde{\sigma_x^2} - \widehat{\sigma_x^2}\right) \xrightarrow{P} 0$ and $\widehat{\sigma_x^2} \xrightarrow{P} c_b \neq 0$ by Lemma 11.

We have already established that $\sqrt{n}(\widetilde{\overline{x^2}} - \overline{x^2}) \xrightarrow{P} 0$ and $\sqrt{n}\left(\sqrt{\widetilde{\sigma_x^2}} - \sqrt{\widehat{\sigma_x^2}}\right) \xrightarrow{P} 0$.

As a result, by Lemma 16,

$$\sqrt{n}\left(\sqrt{\widetilde{\overline{x^2}} + 1 + 2\sqrt{\widetilde{\sigma_x^2}}} - \sqrt{\overline{x^2} + 1 + 2\sqrt{\widehat{\sigma_x^2}}}\right) \xrightarrow{P} 0.$$

Then since $\hat{E}_n, \tilde{E}_n$ converge to constant matrices and $\sqrt{n}(\tilde{\bar{x}} - \bar{x}) \xrightarrow{P} 0$, $\sqrt{n}(\widetilde{\overline{x^2}} - \overline{x^2}) \xrightarrow{P} 0$, $\sqrt{n}\left(\sqrt{\widetilde{\sigma_x^2}} - \sqrt{\widehat{\sigma_x^2}}\right) \xrightarrow{P} 0$, it follows by Lemma 15 that $\sqrt{n}(\tilde{E}_n - \hat{E}_n) \xrightarrow{P} 0$.

As a corollary, $\tilde{E}_n \xrightarrow{P} C^{1/2}$ since $\hat{E}_n \xrightarrow{P} C^{1/2}$ by Lemma 11.

(11): Also, $\tilde{F}_n \xrightarrow{P} (c_y \quad c_{xy})^T$ since $\tilde{\bar{y}} \xrightarrow{P} c_y$ and $\widetilde{\overline{xy}} \xrightarrow{P} c_{xy}$.

(12): Finally, we show that $\tilde{G}_n \xrightarrow{P} c_{y^2} \in \mathbb{R}$: using the assumption that $\mathbb{P}[\exists i \in [n], y_i \notin [-\Delta_n, \Delta_n]] \to 0$, we can obtain that $\mathbb{P}[\exists i \in [n], y_i^2 \notin [0, \Delta_n^2]] \to 0$ so that $\sqrt{n}|\overline{y^2} - \overset{\smile}{\overline{y^2}}| \xrightarrow{P} 0$. Combining with $\frac{\sqrt{n}}{n}|N_5| \xrightarrow{P} 0$, by the triangle inequality, $\sqrt{n}|\overline{y^2} - \widetilde{\overline{y^2}}| \xrightarrow{P} 0$ which implies that $\sqrt{n}|\tilde{G}_n - \hat{G}_n| \xrightarrow{P} 0$. Then by Lemma 11 and Lemma 25, $\tilde{G}_n \xrightarrow{P} c_{y^2}$. ■

Lemma 17 shows that the noise added to the non-DP estimates converges, in probability, to 0 and that the DP estimates of the regression parameters converge, in probability, to the true parameters. Next, we will show the $1/\sqrt{n}$ convergence rates of $\tilde{\beta}^N, \tilde{\beta}$. As a corollary, this implies the consistency of $\tilde{\beta}^N, \tilde{\beta}$.

Lemma 18 *For every sequence of clipping bounds $\Delta = \Delta_n > 0$ and sequence of privacy parameters $\rho = \rho_n > 0$, in Algorithm 2, under the conditions of Theorem 9:*

1. $\sqrt{n}(\tilde{\beta}^N - \hat{\beta}^N) \xrightarrow{P} 0$,
2. $\sqrt{n}(\tilde{\beta} - \hat{\beta}) \xrightarrow{P} 0$.

Proof [Proof of Lemma 18] As previously defined,

$$\tilde{\beta}^N = \begin{pmatrix} \tilde{\bar{y}} \\ 0 \end{pmatrix}, \quad \hat{\beta}^N = \begin{pmatrix} \bar{y} \\ 0 \end{pmatrix}.$$

Then, $\sqrt{n}(\tilde{\beta}^N - \hat{\beta}^N) \xrightarrow{P} 0$ by Lemma 17 since $\sqrt{n}|\tilde{\bar{y}} - \breve{\bar{y}}| \xrightarrow{P} 0$.

We will show that $\sqrt{n}(\tilde{\beta} - \hat{\beta}) \xrightarrow{P} 0$. First, to show that $\sqrt{n}(\tilde{\beta}_1 - \hat{\beta}_1) \xrightarrow{P} 0$, we apply Lemma 15 with $\widetilde{A} = \widetilde{\overline{xy}} - \tilde{\bar{x}}\tilde{\bar{y}}$, $A = \overline{xy} - \bar{x} \cdot \bar{y}$, $\widetilde{B} = \widetilde{\overline{x^2}} - \tilde{\bar{x}}^2 = \widetilde{\sigma_x^2}$, $B = \overline{x^2} - \bar{x}^2 = \widehat{\sigma_x^2}$ and $f(n) = \sqrt{n}$. Then $\tilde{\beta}_1 = \frac{\widetilde{A}}{\widetilde{B}}$ and $\hat{\beta}_1 = \frac{A}{B}$. By Lemma 11, $B = \widehat{\sigma_x^2}$ converges, in probability, to a non-zero constant and $\sqrt{n}(\widetilde{\overline{xy}} - \overline{xy})$, $\sqrt{n}\left(\bar{x} \cdot \bar{y} - \tilde{\bar{x}}\tilde{\bar{y}}\right) \xrightarrow{P} 0$ by Lemma 17. Also, by Lemma 17 and Lemma 11, if we define $\widetilde{B} = \widetilde{\sigma_x^2}, B = \widehat{\sigma_x^2}, \widetilde{A} = \widetilde{\overline{xy}} - \tilde{\bar{x}}\tilde{\bar{y}}$, then $\sqrt{n}(\widetilde{B} - B) \xrightarrow{P} 0$ and $\sqrt{n}(\widetilde{A} - A) \xrightarrow{P} 0$. Then by Lemma 15, $\sqrt{n}(\tilde{\beta}_1 - \hat{\beta}_1) \xrightarrow{P} 0$. By similar arguments, $\sqrt{n}(\hat{\beta}_2 - \tilde{\beta}_2) \xrightarrow{P} 0$ so that $\sqrt{n}(\hat{\beta} - \tilde{\beta}) \xrightarrow{P} 0$.

■

Lemma 18 leads to the following corollary, showing consistency of the DP estimates of β under the null or alternative hypothesis.

Corollary 19 *For every sequence of clipping bounds $\Delta = \Delta_n > 0$ and sequence of privacy parameters $\rho = \rho_n > 0$, in Algorithm 2, under the conditions of Theorem 9:*

1. *Under the null hypothesis:* $\tilde{\beta}^N \xrightarrow{P} \beta$,
2. *Under the alternative hypothesis:* $\tilde{\beta} \xrightarrow{P} \beta$.

Proof [Proof of Corollary 19] By Lemma 11, $\hat{\beta}_1 \xrightarrow{P} \beta_1$ and $\hat{\beta}_2 \xrightarrow{P} \beta_2$. Then, using Lemma 18 and Lemma 25: $\tilde{\beta}_1 \xrightarrow{P} \beta_1$, $\tilde{\beta}_2 \xrightarrow{P} \beta_2$.

Also, by Lemma 11, $\bar{y} \xrightarrow{P} c_y$ so that under the null hypothesis, $\tilde{\beta}^N \xrightarrow{P} \beta$.

■

5.3 Convergence of Differentially Private F-Statistic

We now introduce the continuous mapping theorem, which is especially useful for combining individual convergence results to show, under certain conditions, more complex convergence results. The continuous mapping theorem can be used to map convergent sequences into another convergent sequence via a continuous function.

Theorem 20 (Continuous Mapping Theorem, see (Gut, 2013)) *Let $\{W_n\}$ be a sequence of random vectors and W be a random vector taking values in the same metric space $\mathcal{X}$. Let $\mathcal{Y}$ be a metric space and $g : \mathcal{X} \to \mathcal{Y}$ be a measurable function.*

Define $D_g = \{x : g \text{ is discontinuous at } x\}$. Suppose that $\mathbb{P}[W \in D_g] = 0$. Then:

1. $W_n \xrightarrow{P} W \Rightarrow g(W_n) \xrightarrow{P} g(W)$,
2. $W_n \xrightarrow{D} W \Rightarrow g(W_n) \xrightarrow{D} g(W)$,
3. $W_n \xrightarrow{a.s.} W \Rightarrow g(W_n) \xrightarrow{a.s.} g(W)$.

We now state and prove a helper lemma that will be useful for showing that the numerators and denominators of the DP F-statistic converge to the right distribution.

Lemma 21 *Let A_n, B_n be random vectors such that there exists distribution L where:*

1. $A_n - B_n \xrightarrow{P} 0$,
2. $\|B_n\| \xrightarrow{D} L$ *such that* $\mathbb{P}[\|B_n\| = 0] = 0$.

Then,

$$\|A_n\|^2 \xrightarrow{D} L^2.$$

Proof [Proof of Lemma 21] Consider the unit vector $\frac{B_n}{\|B_n\|}$. Since $A_n - B_n \xrightarrow{P} 0$, we have that by definition of convergence in probability:

$$\frac{B_n}{\|B_n\|}(A_n - B_n) \xrightarrow{P} 0,$$

where $\|B_n\|$ is almost surely never 0.

First, let $W_n = (\|B_n\|, \|B_n\|)$. Then since $\|B_n\| \xrightarrow{D} L$, by the continuous mapping theorem (Theorem 20), if $g((x, y)) = x \cdot y$, then $g(W_n) = \|B_n\|^2 \xrightarrow{D} L^2$.

Then, let $W_n = (\|B_n\|, \frac{B_n}{\|B_n\|}(A_n - B_n))$. Then since $\mathbb{P}[\|B_n\| = 0] = 0$, by the continuous mapping theorem (Theorem 20), if $g((x, y)) = x \cdot y$, then $g(W_n) = B_n \cdot (A_n - B_n) \xrightarrow{D} 0$ which implies that

$$\langle A_n, B_n \rangle - \|B_n\|^2 = B_n \cdot (A_n - B_n) \xrightarrow{P} 0,$$

so that

$$2(\langle A_n, B_n \rangle - \|B_n\|^2) \xrightarrow{P} 0. \tag{15}$$

Also, by the continuous mapping theorem,

$$\|A_n\|^2 - 2\langle A_n, B_n\rangle + \|B_n\|^2 \tag{16}$$
$$= \|A_n - B_n\|^2 \tag{17}$$
$$\xrightarrow{P} 0. \tag{18}$$

Adding Equations (15) and (16) results in the following: $\|A_n\|^2 - \|B_n\|^2 \xrightarrow{P} 0$. Then by Lemma 25, since $\|B_n\|^2 \xrightarrow{D} L^2$, we have that $\|A_n\|^2 \xrightarrow{D} L^2$.

■

We now show that the main terms in the numerators and denominators of the DP F-statistic converge to the asymptotic distribution of their non-private counterparts.

Lemma 22 *Let $\sigma_e > 0$, $r = p = 2, q = 1$, and $\beta \in \mathbb{R}^p$. For every $n \in \mathbb{N}$ with $n > r$, let $X_n \in \mathbb{R}^{n\times p}$ be the design matrix. For every sequence of clipping bounds $\Delta = \Delta_n > 0$ and sequence of privacy parameters $\rho = \rho_n > 0$, in Algorithm 2, under the conditions of Theorem 9:*

$$\frac{n(\tilde{\beta}^T\tilde{E}_n^2\tilde{\beta} - 2\tilde{\beta}^T\tilde{F}_n + \tilde{G}_n)}{n-r} \xrightarrow{P} \sigma_e^2, \quad \|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2 \xrightarrow{D} \mathcal{X}^2_{r-q}(\eta^2)\sigma_e^2.$$

Proof [Proof of Lemma 22] By Lemma 11 and Lemma 17:

1. $\tilde{\beta}, \hat{\beta} \xrightarrow{P} \beta$,
2. $\tilde{E}_n, \hat{E}_n \to C^{1/2}$,
3. $\tilde{F}_n, \hat{F}_n \xrightarrow{P} (c_y \quad c_{xy})^T$,
4. $\tilde{G}_n, \hat{G}_n \xrightarrow{P} c_{y^2}$.

Furthermore,

$$\frac{n}{n-r} = \frac{1}{1 - r/n} \to 1.$$

As a result, by Slutsky's Theorem,

$$\left(\frac{n(\tilde{\beta}^T\tilde{E}_n^2\tilde{\beta} - 2\tilde{\beta}^T\tilde{F}_n + \tilde{G}_n) - n(\hat{\beta}^T\hat{E}_n^2\hat{\beta} - 2\hat{\beta}^T\hat{F}_n + \hat{G}_n)}{n-r}\right) \xrightarrow{P} 0.$$

By Theorem 3,

$$\frac{\|Y_n - X_n\hat{\beta}\|^2}{n-r} \xrightarrow{P} \sigma_e^2.$$

By Lemma 10, $\|Y - X\hat{\beta}\|^2 = n(\hat{\beta}^T\hat{E}_n^2\hat{\beta} - 2\hat{\beta}^T\hat{F}_n + \hat{G}_n)$. As a result, by Lemma 25,

$$\frac{n(\tilde{\beta}^T\tilde{E}_n^2\tilde{\beta} - 2\tilde{\beta}^T\tilde{F}_n + \tilde{G}_n)}{n-r} \xrightarrow{P} \sigma_e^2.$$

Next, by Theorem 3, $\|X_n\hat{\beta} - X_n\hat{\beta}^N\|^2 \sim \chi^2_{r-q}(\eta^2)\sigma_e^2$. Then by Lemma 10, $\|\sqrt{n}\hat{E}_n(\hat{\beta} - \hat{\beta}^N)\|^2 \sim \chi^2_{r-q}(\eta^2)\sigma_e^2$. We will show that

$$\|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2 \xrightarrow{D} \chi^2_{r-q}(\eta^2)\sigma_e^2.$$

First, we define the random vectors

$$\tilde{H}_n = \tilde{E}_n\sqrt{n}(\tilde{\beta} - \tilde{\beta}^N), \quad \hat{H}_n = \hat{E}_n\sqrt{n}(\hat{\beta} - \hat{\beta}^N).$$

By Lemma 18, $\sqrt{n}(\tilde{\beta} - \hat{\beta}) \xrightarrow{P} 0$ and $\sqrt{n}(\tilde{\beta}^N - \hat{\beta}^N) \xrightarrow{P} 0$. And since by Lemma 11, $\hat{E}_n \xrightarrow{P} C^{1/2}$, we have that by Slutsky's Theorem, $\sqrt{n}\hat{E}_n(\tilde{\beta} - \hat{\beta}) \xrightarrow{P} 0$, and $\sqrt{n}\hat{E}_n(\tilde{\beta}^N - \hat{\beta}^N) \xrightarrow{P} 0$. Also, by Lemma 17, $\sqrt{n}(\tilde{E}_n - \hat{E}_n) \xrightarrow{P} 0$ which implies that, by Slutsky's Theorem, and Lemma 18, $\sqrt{n}(\tilde{E}_n - \hat{E}_n)\tilde{\beta} \xrightarrow{P} 0$ and $\sqrt{n}(\tilde{E}_n - \hat{E}_n)\tilde{\beta}^N \xrightarrow{P} 0$ so that $\sqrt{n}(\tilde{E}_n - \hat{E}_n)(\tilde{\beta} - \tilde{\beta}^N) \xrightarrow{P} 0$.

As a result,

$$\begin{aligned}
&\tilde{H}_n - \hat{H}_n \\
&= \tilde{E}_n\sqrt{n}(\tilde{\beta} - \tilde{\beta}^N) - \hat{E}_n\sqrt{n}(\hat{\beta} - \hat{\beta}^N) \\
&= \tilde{E}_n\sqrt{n}\tilde{\beta} - \hat{E}_n\sqrt{n}\hat{\beta} - \left[\tilde{E}_n\sqrt{n}\tilde{\beta}^N - \hat{E}_n\sqrt{n}\hat{\beta}^N\right] \\
&= (\tilde{E}_n - \hat{E}_n)\sqrt{n}\tilde{\beta} + \hat{E}_n\sqrt{n}(\tilde{\beta} - \hat{\beta}) - \left[(\tilde{E}_n - \hat{E}_n)\sqrt{n}\tilde{\beta}^N + \hat{E}_n\sqrt{n}(\tilde{\beta}^N - \hat{\beta}^N)\right] \\
&= (\tilde{E}_n - \hat{E}_n)\sqrt{n}(\tilde{\beta} - \tilde{\beta}^N) + \hat{E}_n\sqrt{n}(\tilde{\beta} - \hat{\beta}) - \left[\hat{E}_n\sqrt{n}(\tilde{\beta}^N - \hat{\beta}^N)\right] \\
&\xrightarrow{P} 0.
\end{aligned}$$

By Lemma 21, since $\|\hat{H}_n\| \xrightarrow{D} \chi_{r-q}(\eta^2)\sigma_e$ and $\tilde{H}_n - \hat{H}_n \xrightarrow{P} 0$, we have that

$$\|\tilde{H}_n\|^2 \xrightarrow{D} \chi^2_{r-q}(\eta^2)\sigma_e^2.$$

As a result, $\|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2 \xrightarrow{D} \chi^2_{r-q}(\eta^2)\sigma_e^2$. This completes the proof.

■

Lemma 22 shows the convergence of individual quantities that can now be combined to show the convergence of the DP test statistic $\tilde{T}$:

Proof [Proof of Theorem 9] First, by Corollary 19, under the null hypothesis: $\tilde{\beta}^N = \tilde{\beta}_n^N \xrightarrow{P} \beta$. And under the alternative hypothesis: $\tilde{\beta} = \tilde{\beta}_n \xrightarrow{P} \beta$.

By Lemma 10,

$$T = T_n = \frac{n-r}{r-q} \cdot \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2} = \frac{n-r}{r-q} \cdot \frac{\|\sqrt{n}\hat{E}_n(\hat{\beta} - \hat{\beta}^N)\|^2}{n(\hat{\beta}^T\hat{E}_n^2\hat{\beta} - 2\hat{\beta}^T\hat{F}_n + \hat{G}_n)}.$$

And by Equation (12),

$$\tilde{T} = \tilde{T}_n = \frac{n-r}{r-q} \cdot \frac{\|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2}{n(\tilde{\beta}^T\tilde{E}_n^2\tilde{\beta} - 2\tilde{\beta}^T\tilde{F}_n + \tilde{G}_n)}.$$

From Theorem 3, in the non-private case where $Y_n \sim \mathcal{N}(X_n\beta, \sigma_e^2 I_{n\times n})$, if $T = T_n$ is the test statistic from Equation (4), then

$$T_n \sim F_{r-q,n-r}(\eta_n^2), \quad \eta_n^2 = \frac{\|X_n\beta - X_n\beta^N\|^2}{\sigma_e^2}.$$

Also, by Theorem 3, the asymptotic distribution of T is a chi-squared distribution. i.e., $T = T_n \xrightarrow{D} \frac{\chi^2_{r-q}(\eta^2)}{r-q}$. Next, we show that the DP F-statistic also has asymptotic distribution of chi-squared.

By Lemma 22,

$$\|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2 \xrightarrow{D} \mathcal{X}^2_{r-q}(\eta^2)\sigma_e^2,$$

and

$$\frac{n(\tilde{\beta}^T\tilde{E}_n^2\tilde{\beta} - 2\tilde{\beta}^T\tilde{F}_n + \tilde{G}_n)}{n-r} \xrightarrow{P} \sigma_e^2.$$

Let

$$W_n = \left(\frac{\|\sqrt{n}\tilde{E}_n(\tilde{\beta} - \tilde{\beta}^N)\|^2}{r-q}, \frac{n(\tilde{\beta}^T\tilde{E}_n^2\tilde{\beta} - 2\tilde{\beta}^T\tilde{F}_n + \tilde{G}_n)}{n-r}\right) = (A_n, B_n),$$

and

$$W = \left(\frac{\mathcal{X}^2_{r-q}(\eta^2)\sigma_e^2}{r-q}, \sigma_e^2\right).$$

By the condition that $\sigma_e > 0\,(\mathbb{P}[\sigma_e = 0] = 0)$ and the continuous mapping theorem (Theorem 20), $\tilde{T}_n = g(W_n) = g(A_n, B_n) = A_n/B_n$ converges, in distribution, to $\frac{\chi^2_{r-q}(\eta^2)}{r-q}$ as $n \to \infty$.

■

Lemma 23 (Weak Law of Large Numbers, see (Keener, 2010)) *Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with mean μ. Then*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n \xrightarrow{P} \mu,$$

provided that $\mathbb{E}[|Y_i|] < \infty$.

Lemma 24 *Let $X \sim F_{n,m}(\lambda)$ and $Y = \lim_{m\to\infty} nX$. Then $Y \sim \chi^2_n(\lambda)$.*

A self-contained proof of Lemma 24 can be found in (Alabi and Vadhan, 2022).

Lemma 25 *Let $\{X_n\}$ and $\{Y_n\}$ be a sequence of random vectors and X be a random vector. Then:*

1. *If $X_n \xrightarrow{D} X$ and $X_n - Y_n \xrightarrow{P} 0$, then $Y_n \xrightarrow{D} X$.*

2. *If* $X_n \xrightarrow{P} X$*, then* $X_n \xrightarrow{D} X$*.*

3. *For a constant* $c \in \mathbb{R}$*, if* $X_n \xrightarrow{D} c$*, then* $X_n \xrightarrow{P} c$*.*

Proof Follows from Theorem 2.7 in (Vaart, 1998). ■

6. Experimental Evaluation

We will measure the effectiveness of our hypothesis tests via significance and power. The power and significance of our differentially private tests are estimated on both semi-synthetic datasets based on the Opportunity Atlas (Chetty et al., 2018, 2019) and on synthetic datasets. The OI semi-synthetic datasets consists of simulated microdata for each census tract in some states in the U.S. The dependent variable Y is the child national income percentile and the independent variable X is the corresponding parent national income percentile. See (Alabi et al., 2022) for more details on the properties of simulated data from the OI team. In the OI data, X is lognormally distributed and the distribution of counts of individuals across tracts in a state follows an exponential distribution.

6.1 General Parameter Setup for Synthetic Data

For experimental evaluation on synthetic datasets, we generated datasets with sizes between $n = 100$ and $n = 10,000$.

For both the linear relationship and mixture model tests on synthetic data below, we consider a subset of the following values of the privacy budget ρ: $\{0.1^2/2, 0.5^2/2, 1^2/2, 2^2/2, 3^2/2, 5^2/2, 10^2/2\}$.

We draw the independent variables $x_1, \ldots, x_n$ according to a few different distributions: Normal, Uniform, Exponential. We will detail the parameters used to generated variables from these distributions in the corresponding subsections.

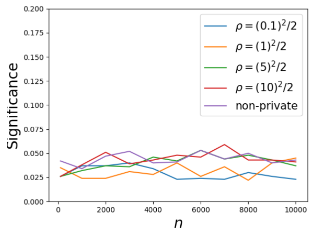
For all tests below, the clipping parameter is either set to $\Delta = 2$ or $\Delta = 3$. For the synthetic data, the dependent variable Y is generated using the linear or mixture model specification described in previous sections and by fixing or varying parameters (such as σ_e). For estimating the power and significance, we fix the target significance level to 0.05 and run Monte Carlo tests 2000 times. We estimate the power and significance as the fraction of times the null is rejected, given various settings of parameters that satisfy the alternative and null hypothesis, respectively.

6.2 Testing a Linear Relationship on Synthetic Data

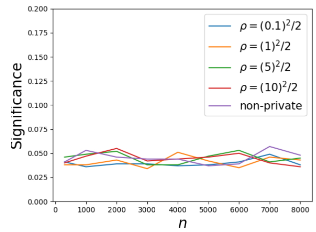
6.2.1 F-statistic

We evaluate our DP linear relationship test on synthetically generated data from three different distributions: normal, uniform, and exponential. We also vary parameters such as: the slope of the linear model and the noise distribution of the dependent variable.

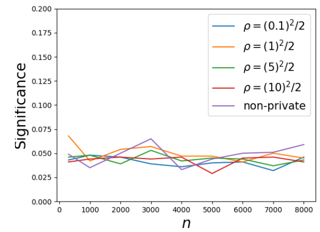
Evaluating the Significance for Normally Distributed Independent Variables: Generally, we see that the significance remains below the target significance level, on average, for all values of ρ. For the linear relationship tester, when the standard deviation of the



(a) Significance for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 0.001^2)$. $\Delta = 2$.

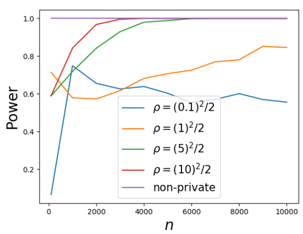


(b) Significance for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

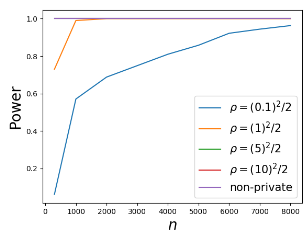


(c) Significance for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 1)$. $\Delta = 2$.

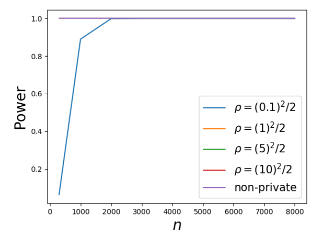
Figure 1



(a) Power for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.001^2)$. $\Delta = 2$.



(b) Power for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.



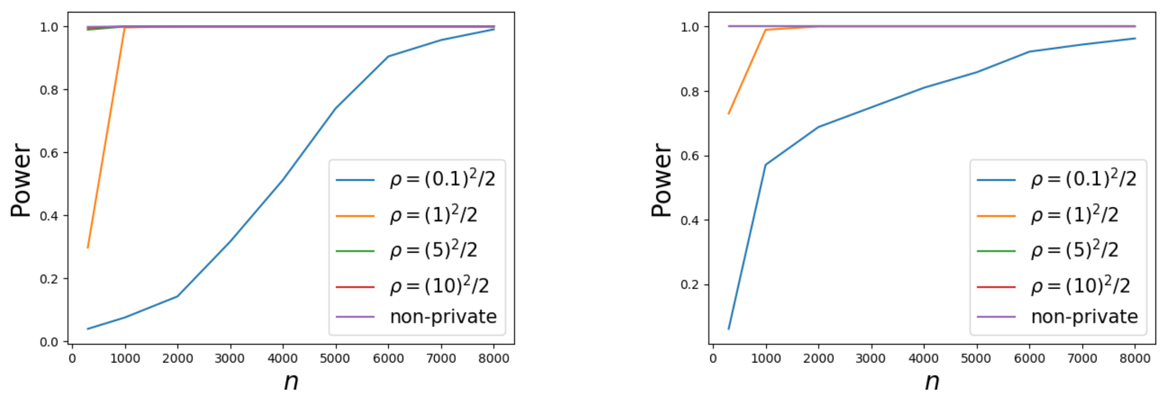
(c) Power for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$. $\Delta = 2$.

Figure 2

dependent variable (σ_e) is small (Figure 1a), we see that the true significance level is well below the target significance of 0.05, which is fine (but conservative). We conjecture that this happens because when σ_e is small: (i) we fail to reject when the noisy estimate of σ_e is ≤ 0; or (ii) the test statistic under the null distribution will be almost always 0 since under the null (even without privacy), the standard deviation of the test statistic is proportional to σ_e. In Figures 1a, 1b and 1c, we see the significance of the linear tester attains the target (of 0.05) as we vary the noise in the dependent variable σ_e.

Evaluating Power for Varying the Noise in the Dependent Variable: For Figures 2a, 2b, and 2c, we set the true slope to 1. We then vary the noise in the dependent variable. That is, for the multivariate linear regression model, $Y \sim \mathcal{N}(X\beta, \sigma_e^2 I_{n \times n})$, we vary σ_e. The following values of σ_e are considered: $\{0.001, 0.35, 1\}$.

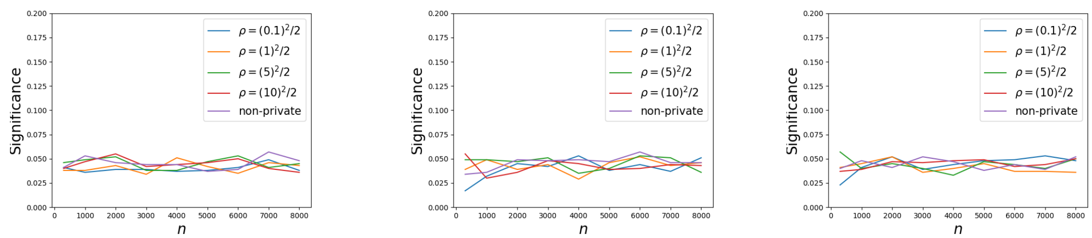
In Figure 2a, we generally see that compared to higher values of σ_e (Figures 2b and 2c), the power is relatively low. We believe this occurs because when σ_e is small, its DP estimate is more likely to be less than 0, in which case we fail to reject the null (even when the alternative is true). This generally leads to a reduction in the power.



(a) Power for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0.1 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

(b) Power for testing a linear relationship. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

Figure 3



(a) Significance for testing a linear relationship. Normal Distribution on X.

(b) Significance for testing a linear relationship. Uniform Distribution on X.

(c) Significance for testing a linear relationship. Exponential Distribution on X.

Figure 4

Evaluating Power for Varying Slopes: Figures 3a, 3b show the power of the linear test for slopes of 0.1, 1. We generally see that the larger the slope, the higher the power of the DP tests.

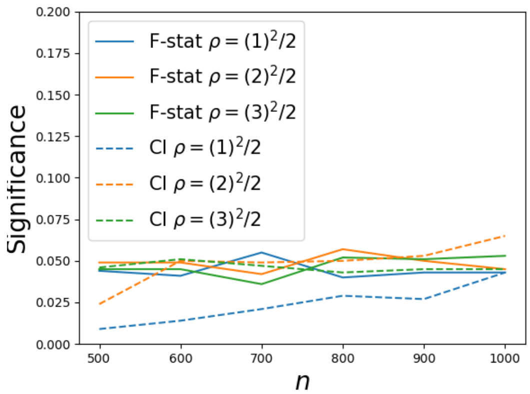
Evaluating the Significance while Varying the Distribution of the Independent Variable: For Figures 4a, 4b, and 4c, we set the standard deviation of the noise dependent variable to 0.35. We then vary the distribution of the independent variable — while maintaining the variance — to take on one of the following:

1. **Normal**: with mean 0.5 and variance 1/12.

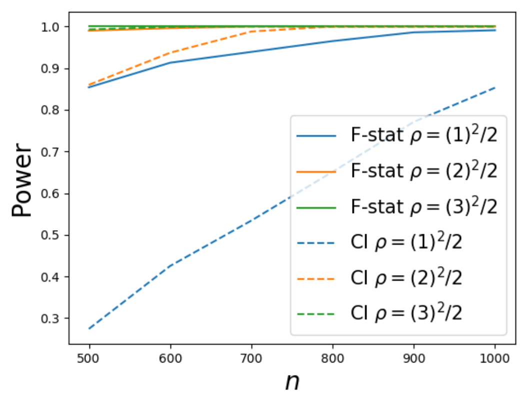
2. **Uniform**: between 0 and 1 (variance of 1/12).

3. **Exponential**: with scale of $1/\sqrt{12}$.

We observe that the significance is still preserved even though, in our DP testers, the null distribution is simulated via a normal distribution.



(a) Significance for F-statistic versus confidence interval approach. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.



(b) Power for F-statistic versus confidence interval approach. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

Figure 5

6.2.2 Bootstrap Confidence Intervals

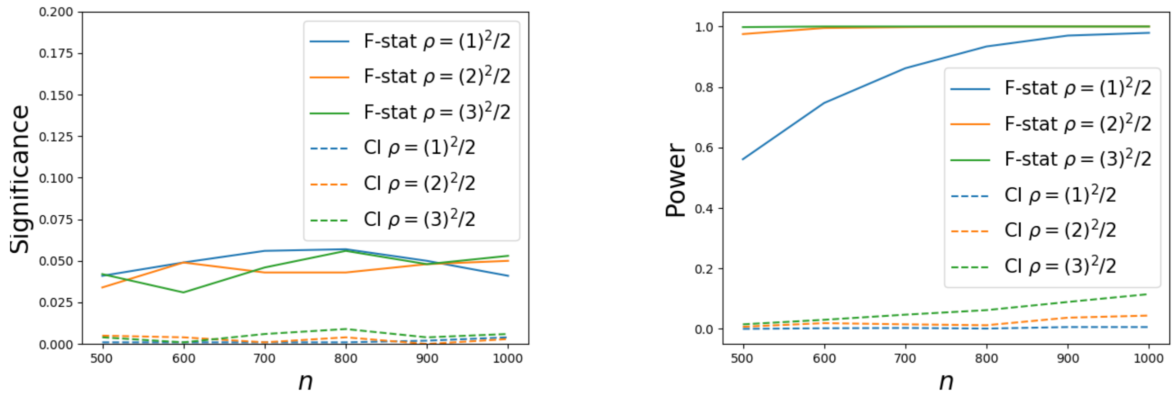
Using the duality between confidence interval estimation and hypothesis testing, we can construct hypothesis tests for testing a linear relationship based on DP confidence interval procedures. Specifically, we compare the F-statistic linear relationship tester to the tester that uses DP confidence intervals (DP CI).

In Figure 5a, we present experimental results for the significance level of DP CI compared to Algorithm 1 instantiated with the DP F-statistic. As we see, DP CI achieves the target significance level. In Figure 5b, we also present experimental results for the power of DP CI compared to Algorithm 1. We see that DP CI has less power than Algorithm 1. This observation is more pronounced for less concentrated distributions (i.e., uniform) on the independent variable. See Figure 6b. This might be due to the, sometimes excessive, width of the confidence interval produced by the bootstrap interval (in order to ensure coverage under the null hypothesis).

Figures 5a, 5b, 6a, and 6b show results averaged out over 2000 trials. The dashed lines correspond to the bootstrap confidence interval approach (denoted CI) while the solid lines are for the F-statistic (denoted F-stat).

6.2.3 Bernoulli Tester

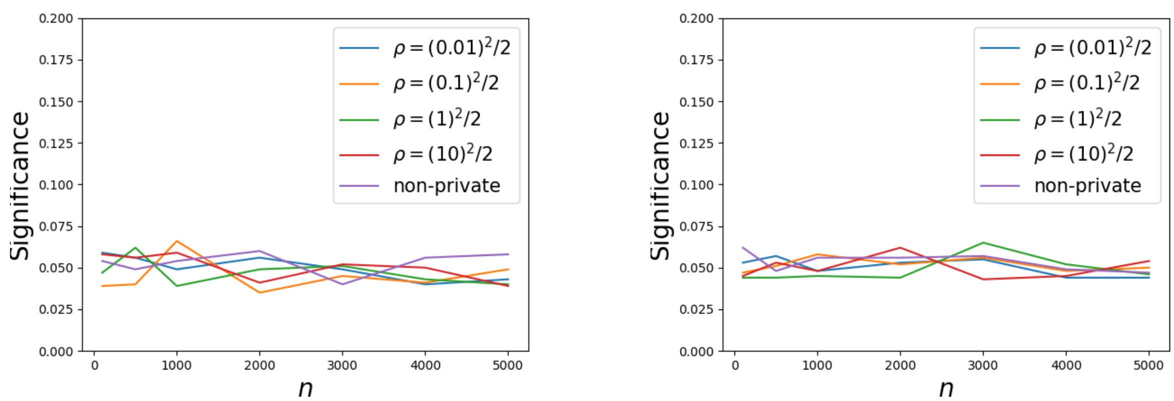
The DP Bernoulli tester has a higher power than the DP F-statistic when the slope is large or the privacy-loss parameter is small; otherwise the DP F-statistic performs better. We vary the slope from 0.01 up to 1. As the slope gets smaller, the performance gap between the F-statistic and the Bernoulli tester gets larger. Figures 8a and 8b show the power of the DP F-statistic compared to the DP Bernoulli tester for the problem of testing a linear relationship. The significance levels of the Bernoulli tester are shown in Figures 7a and 7b.



(a) Significance for F-statistic versus confidence interval approach. $x_i \sim \text{Unif}[0,1]$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

(b) Power for F-statistic versus confidence interval approach. $x_i \sim \text{Unif}[0,1]$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

Figure 6

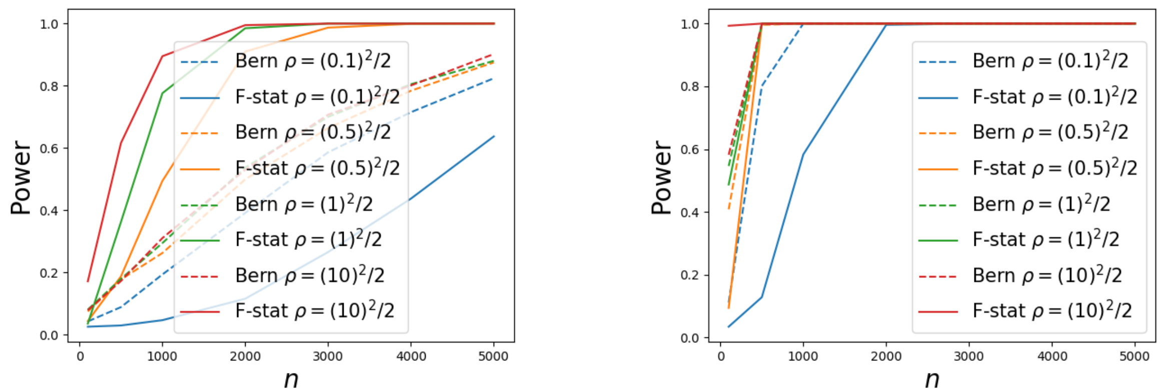


(a) Significance for testing for a fair coin. n observations are generated from Bern(1/2).

(b) Significance for linear relationship testing via Bernoulli testing approach. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 1)$. $\Delta = 2$.

Figure 7

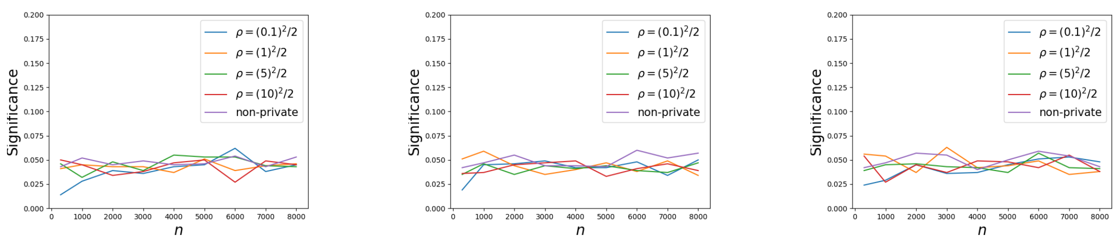
The results are averaged out over 1000 trials. For the figures illustrating the power, dashed lines correspond to the Bernoulli testing approach (denoted Bern) while the solid lines are for the F-statistic (denoted F-stat).



(a) Power for F-statistic versus Bernoulli testing approach. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0.1 \cdot x_i + \mathcal{N}(0, 1)$. $\Delta = 2$.

(b) Power for F-statistic versus Bernoulli testing approach. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0.5 \cdot x_i + \mathcal{N}(0, 1)$. $\Delta = 2$.

Figure 8



(a) Significance for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.01^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.01^2)$ for Group 2.

(b) Significance for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.

(c) Significance for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 2.

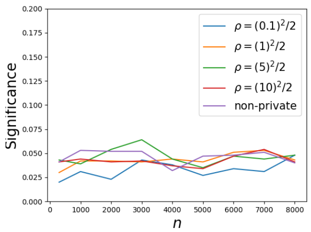
Figure 9

6.3 Testing Mixture Models on Synthetic Data

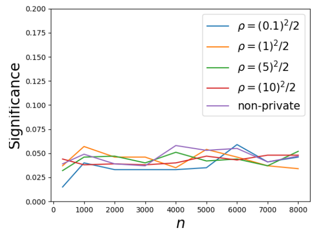
6.3.1 F-Statistic

We evaluate the F-statistic DP mixture model test on synthetically generated data. We vary parameters such as: the fraction of data in each group and the slopes used to generate data for each group. Let β_1, β_2 denote the slopes of the two groups.

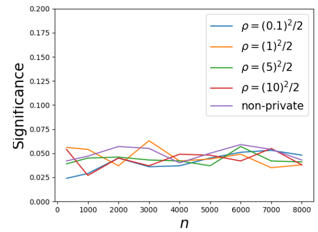
Evaluating the Significance: Like in the DP linear model tester, we also see that we achieve the target significance levels, on average, for all values of ρ. In Figures 9a, 9b, and 9c, we vary the noise in the dependent variable. In Figures 10a, 10b, and 10c, we vary the fraction of group sizes, using either a 1/8, 1/4, or 1/2 fraction for the first group. We see that the more unbalanced splits tend to have lower significance levels.



(a) Significance for testing mixtures. 1/8th vs. 7/8th splits. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.

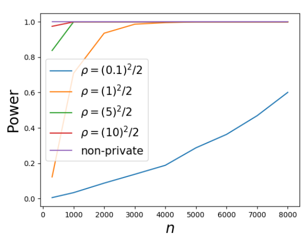


(b) Significance for testing mixtures. 1/4th vs. 3/4th splits. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.

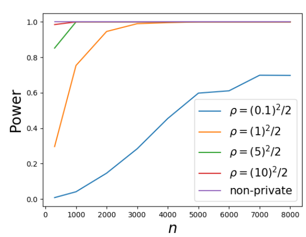


(c) Significance for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.

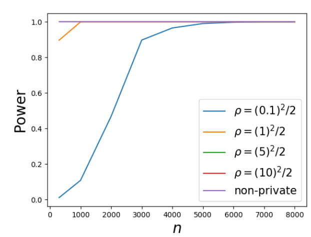
Figure 10



(a) Power for testing mixtures. 1/8th vs. 7/8th splits. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.



(b) Power for testing mixtures. 1/4th vs. 3/4th splits. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.



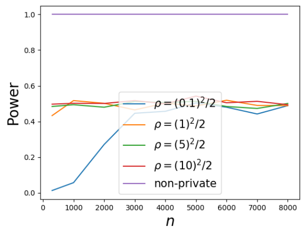
(c) Power for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 2.

Figure 11

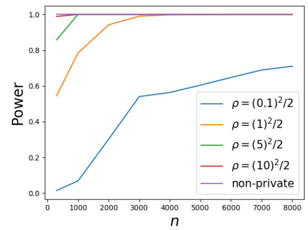
Power while Varying the Group Size Fraction: Let n be the total number of datapoints and n_1, n_2 be the number of points in groups 1 and 2 respectively. We vary the fraction of points in group 1: n_1/n. Setting the slopes of each group to $\beta_1 = -1$ and $\beta_2 = 1$, we vary this fraction so that $n_1/n \in \{1/8, 1/4, 1/2\}$. For Figure 11a, we set the group sizes to be equal. For Figure 11b, we set the group sizes to be $n/4, 3n/4$. And last, for Figure 11c, the group sizes are $n/8, 7n/8$.

Generally, the more even the group size fractions are, the higher the power of the DP test for testing mixtures in the multivariate linear regression model.

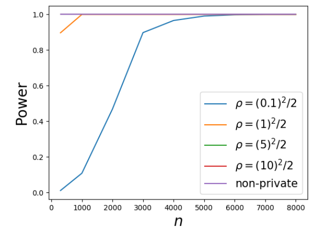
Power while Varying the Noise in the Dependent Variable: We also vary σ_e. We generally see that the smaller it is, the smaller the power. We conjecture that this happens because we err on the side of failing to reject the null if the DP estimate of σ_e becomes ≤ 0, which is more likely to happen if σ_e is small. In Figures 12a, 12b,and 12c we see this phenomenon.



(a) Power for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 0.01^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.01^2)$ for Group 2.

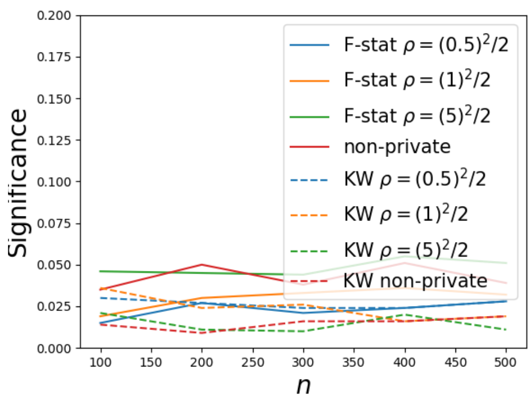


(b) Power for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.

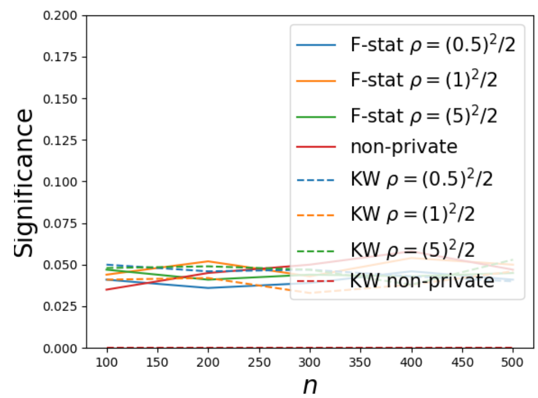


(c) Power for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 2.

Figure 12



(a) Significance for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 0.1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$ for Group 2.



(b) Significance for testing mixtures. Equal-sized groups. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 2.

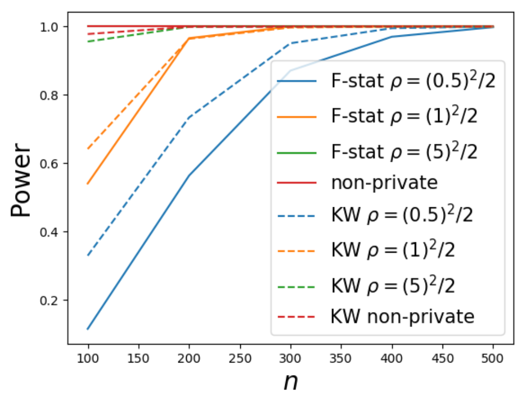
Figure 13

6.3.2 Nonparametric Tests via Kruskal-Wallis

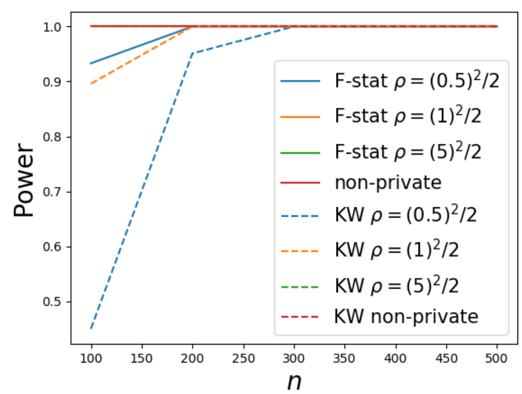
We now proceed to show results for comparing the mixture models based on Kruskal-Wallis (KW) to the parametric F-statistic method.

Evaluating the Significance: The KW methods, on average, achieve the target significance levels for all values of ρ as illustrated in Figures 13a and 13b, where we vary the noise in the dependent variable.

Evaluating the Power as we Increase the Variance of the Independent Variable: In Figures 14a and 14b, we see that the F-statistic method outperforms the KW method when the variance of the independent variable is much larger (10x) than previously.



(a) Power for Kruskal-Wallis versus the F-statistic. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 2.



(b) Power for Kruskal-Wallis versus the F-statistic. $x_i \sim \mathcal{N}(0.5, 10)$, $y_i \sim -1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 1. $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 1)$ for Group 2.

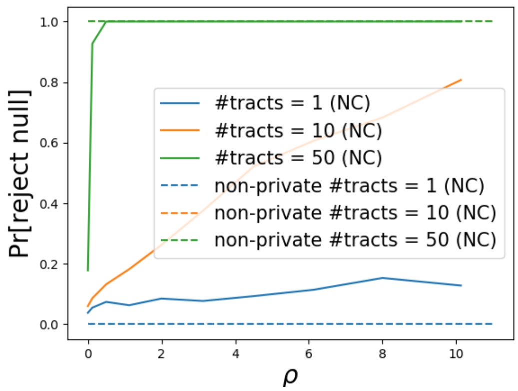
Figure 14

6.4 Testing on Opportunity Insights Data

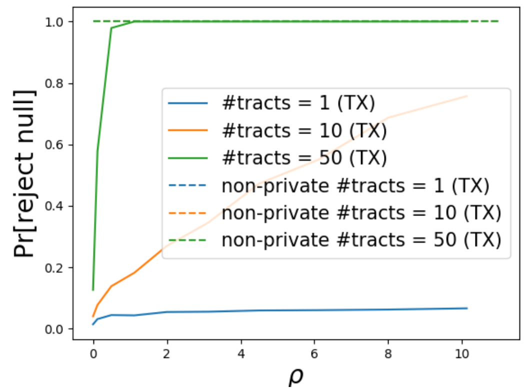
The Opportunity Insights (OI) team gave us simulated data for census tracts from the following states in the United States: Idaho, Illinois, New York, North Carolina, Texas, and Tennessee. The dependent and independent variables are the child and parent national income percentiles, respectively. For the linear tester, a rejection of the null hypothesis implies that there is a relationship between the parent and child income percentiles. For the mixture model tester, it implies that there is more than one linear relationship in the data which suggests that more granular data is needed for analysis on the data. The groups of data fed to the mixture model tester are conglomeration of one or more tracts.

Some of these states have a small number of datapoints. For example, within Illinois, there are tracts with just $n = 39$ datapoints. For the Illinois dataset, there are $n = 219,594$ datapoints that are subdivided into $3,108$ census tracts. The North Carolina and Texas datasets consists of datapoints subdivided into $2,156$ and $5,187$ census tracts respectively. We will focus on data from North Carolina (NC), and Texas (TX) and experimentally evaluate $\mathbb{P}$[reject null], the probability of rejecting the null hypothesis over the randomness of the DP algorithms. We run our tests on some census tracts in these states showing how these measures fair as the privacy parameter is relaxed. For the experiments below, from each state, we randomly and uniformly select: (i) a single tract; (ii) 10 randomly selected tracts and concatenate; and (iii) 50 randomly selected tracts and concatenate. Then we test for the presence of a (non-zero) linear relationship. The concatenation could result in hundreds or thousands of points.

Our tests are evaluated on the OI data. We have not included the test based on Kruskal-Wallis as our current implementation is, at the moment, relatively computationally inefficient to evaluate on such large datasets. See above synthetic data experiments for comparison of Kruskal-Wallis to the F-statistic method. Figures 15a, and 15b show the probability of rejecting the null as we increase the parameter ρ when using the DP linear tester. Fig-

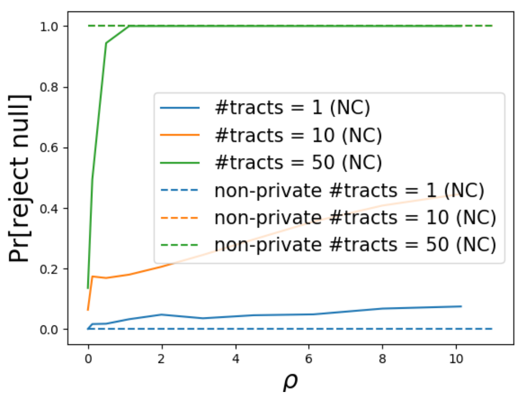


(a) $\mathbb{P}$[reject null] for testing a linear relationship in NC. $\Delta = 2$.

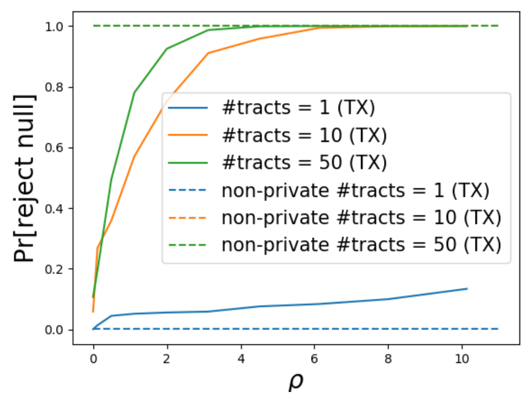


(b) $\mathbb{P}$[reject null] for testing a linear relationship in TX. $\Delta = 2$.

Figure 15



(a) $\mathbb{P}$[reject null] for testing for mixtures in NC. $\Delta = 2$.



(b) $\mathbb{P}$[reject null] for testing for mixtures in TX. $\Delta = 2$.

Figure 16

ures 16a, and 16b show the corresponding results for the F-stat based DP mixture model tester. We see that for the small-sized datasets tend to have a small chance of rejecting the null while larger ones have a higher chance.

6.5 Testing on UCI Bike Dataset

We use the UCI bike dataset (Fanaee-T and Gama, 2014) with 17,389 instances. For this dataset, we test for a linear relationship between the "temp" (normalized temperature in Celsius) and "hr" (hour between 0 and 23) attributes. The null hypothesis is that there is no linear relationship between the "temp" and "hr" attributes. Without privacy, the linear relationship tester based on the F-statistic rejects the null. In Table 1, we show

ρ	0.005	0.125	0.5	1.125	2.0	3.125	4.5	6.125	8.0	10.125	non-DP
$\mathbb{P}$[reject null | 100% data]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\mathbb{P}$[reject null | 10% data]	0.85	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 1: $\mathbb{P}$[reject null] for testing for a linear relationship between temperature and time (in hours).

the probability of Algorithm 1 rejecting the null as we vary the privacy parameter. We can observe that for almost all—except for the smallest setting of ρ—privacy parameters, $\mathbb{P}$[reject null | p% data] (probability of rejecting the null, given p% of the dataset) for the private test matches that of the non-private test.

While we show that our methods can run on real-world datasets, the synthetically generated datasets give a lot more information on the behavior of the tests.

6.6 More Details on Experimental Evaluation of DP Confidence Intervals

We now proceed to construct a hypothesis test based on DP parametric bootstrap confidence intervals (e.g., using the work of (Ferrando et al., 2021)). Then we will experimentally compare to our linear relationship tester based on the DP F-statistic.

Suppose that θ is the set of parameters (e.g., standard deviation of the dependent and independent variables) and $f = f(\theta)$ is the estimation target (e.g., the slope in the dataset). The goal is to obtain a $1-\alpha$ confidence interval $[\hat{a}_n, \hat{b}_n]$ for $f(\theta)$ via an end-to-end differentially private procedure. In other words, we want

$$\mathbb{P}\left[\hat{a}_n \leq f(\theta) \leq \hat{b}_n\right] = 1 - \alpha,$$

where the probability is taken over both θ and f.

Because of the randomized nature of (non-trivial) DP procedures, the finite-sample coverage of the interval might not exactly be close to $1 - \alpha$. Ferrando, Wang, and Sheldon (Ferrando et al., 2021) show the consistency of these intervals (in the large-sample, asymptotic regime).

Algorithm 6 follows the same framework as Algorithm 1, except that instead of simulating test statistics under the null hypothesis, the goal is to calculate a confidence interval for the slope. $P_{(\tilde{\theta}_0, \tilde{\theta}_1)}$ denotes the distribution from which we shall generate our bootstrap samples and from which a confidence interval can be estimated. For example, for taking bootstrap samples for the slope, $P_{(\tilde{\theta}_0, \tilde{\theta}_1)}$ would approximately be distributed as $\mathcal{N}(\tilde{\beta}_1, \frac{\widetilde{S^2}}{\widetilde{\text{nvar}}})$ where $\widetilde{\text{nvar}} = n \cdot \widetilde{x^2} - n \cdot \tilde{x}^2$ and $\widetilde{S^2}$ is as defined in Algorithm 2. Note that a crucial difference between tests based on the parametric bootstrap confidence intervals and our tests is the following: our tests only use $\tilde{\theta}_0$, a subset of the estimated DP statistics, to simulate the null distribution and decide to reject the null while the other approach uses $(\tilde{\theta}_0, \tilde{\theta}_1)$ to decide to reject the null.

The target slope is b. For example, if we seek to test for a linear relationship, we set $b = 0$ since under the null hypothesis, the slope will be 0. `DPStats` is a ρ-zCDP procedure for estimating DP sufficient statistics for a parametric model. In Algorithm 6, $(s_{(l)}, s_{(r)})$ is the parametric bootstrap confidence interval for the slopes under the null hypothesis.

Algorithm 6: DP Test Framework via Parametric Bootstrap Confidence Intervals.

Data: $X \in \mathbb{R}^{n \times p}; Y \in \mathbb{R}^n$
Input: n (dataset size); ρ (privacy-loss parameter); α (target significance); b (target slope)
$(\tilde{\theta}_0, \tilde{\theta}_1) = \texttt{DPStats}(X, Y, n, \rho)$
if $\tilde{\theta}_0 = \tilde{\theta}_1 = \perp$ **then**
  return Fail to Reject the null

Select $K > 1/\alpha$
for $k = 1 \dots K$ **do**
  Sample slope $s_k \sim P_{(\tilde{\theta}_0, \tilde{\theta}_1)}$

Sort $s_{(1)} \leq \cdots \leq s_{(K)}$

Set $l = \lceil (K+1)(\alpha/2) \rceil$
Set $r = \lceil (K+1)(1-\alpha/2) \rceil$
if $b \notin (s_{(l)}, s_{(r)})$ **then**
  return Reject the null
else
  return Fail to Reject the null

Without privacy, by Lemma 11, under the null hypothesis, we know that the slopes will be distributed as the following distribution: $\hat{\beta}_1 \sim \mathcal{N}\left(\beta, \frac{\sigma_e^2}{n \cdot \widehat{\sigma_x^2}}\right) \sim \mathcal{N}\left(0, \frac{\sigma_e^2}{n \cdot \widehat{\sigma_x^2}}\right)$. Even as n increases and as we take fresh samples of $\hat{\beta}_1$, the parametric bootstrap confidence interval around $\hat{\beta}_1$ gets smaller and more concentrated around the true value 0. We expect to observe similar behavior when applying DP.

7. Conclusion

We have developed differentially private hypothesis tests for testing a linear relationship in data and for testing for mixtures in linear regression models. We also show that the DP F-statistic converges to the asymptotic distribution of the non-private F-statistic. Through experiments, we show that our Monte Carlo tests achieve significance that is less than the target significance level across a wide variety of experiments. Furthermore, our tests generally have a high power, getting higher as we increase the dataset size and/or relax the privacy parameter. Even on small datasets (in the hundreds) with small slopes, our tests retain the significance level while having a high enough power. Experimental evaluation is done on simulated data for the Opportunity Atlas tool, UCI datasets, and on synthetic datasets of varying distributions on the independent variable (normal, exponential, and uniform). We also identify regimes where the power of our tests are low (e.g., when the variance of the dependent variable is very small).

We have provided formal statements for the DP F-statistic in the asymptotic regime. In particular, this is a first major step for theoretical and practical usability of the DP F-statistic for determining scientific conclusions. We leave to future work the task of theoretically analyzing the procedures in other regimes (e.g., the non-asymptotic regime).

We make the following recommendation to practitioners: in the cases where the dataset sizes are small and the private sufficient statistics might not be reused (for other tasks), we recommend the use of the DP nonparameteric methods for linear regression testing. In the case where the dataset size is large enough and sufficient statistics might be reused (e.g., to calculate the regression parameters itself or perform other statistical tests), we recommend the DP tests based on the F-statistic. Also, the procedures based on the DP F-statistic might be easier to use and explain than the nonparametric methods.

Acknowledgments

D.A. was supported by a Fellowship from Meta AI and Cooperative Agreement CB20ADR0160001 with the U.S. Census Bureau. Work done while a Ph.D. student at Harvard University. S.V. was supported by Cooperative Agreement CB20ADR0160001 and a Simons Investigator Award. The views expressed in this paper are those of the authors and not those of the U.S. Census Bureau or any other sponsor.

References

J. Acharya, Z. Sun, and H. Zhang. Differentially private testing of identity and closeness of discrete distributions. In *NeurIPS*, pages 6879–6891, 2018.

J. Acharya, C. L. Canonne, C. Freitag, and H. Tyagi. Test without trust: Optimal locally private distribution testing. In *AISTATS*, pages 2067–2076, 2019.

D. Alabi and S. Vadhan. Hypothesis testing for differentially private linear regression. In *Advances in Neural Information Processing Systems*, volume 35, pages 14196–14209, 2022.

D. Alabi, A. McMillan, J. Sarathy, A. D. Smith, and S. P. Vadhan. Differentially private simple linear regression. *Proc. Priv. Enhancing Technol.*, 2022(2):184–204, 2022. doi: 10.2478/popets-2022-0041. URL `https://doi.org/10.2478/popets-2022-0041`.

M. Aliakbarpour, I. Diakonikolas, and R. Rubinfeld. Differentially private identity and equivalence testing of discrete distributions. In *ICML*, pages 169–178, 2018.

M. Aliakbarpour, I. Diakonikolas, D. Kane, and R. Rubinfeld. Private testing of distributions via sample permutations. In *NeurIPS*, pages 10877–10888, 2019.

I. Andrews, T. Kitagawa, and A. McCloskey. Inference on winners. Working Paper 25456, National Bureau of Economic Research, January 2019. URL `http://www.nber.org/papers/w25456`.

M. Avella-Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 0(0):1–15, 2020. doi: 10.1080/01621459.2019.1700130. URL `https://doi.org/10.1080/01621459.2019.1700130`.

J. A. Awan and A. Slavkovic. Differentially private inference for binomial data. *Journal of Privacy and Confidentiality*, 10(1), Jan. 2020. doi: 10.29012/jpc.725. URL `https://journalprivacyconfidentiality.org/index.php/jpc/article/view/725`.

A. F. Barrientos, J. Reiter, A. Machanavajjhala, and Y. Chen. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28:440 – 453, 2017.

G. Bernstein and D. R. Sheldon. Differentially private bayesian linear regression. In *NeurIPS*, pages 523–533, 2019.

P. Bleninger, J. Drechsler, and G. Ronning. Remote data access and the risk of disclosure from linear regression: An empirical study. In *PSD*, volume 6344 of *Lecture Notes in Computer Science*, pages 220–233. Springer, 2010.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC*, pages 635–658, 2016.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825 – 2850, 2021. doi: 10.1214/21-AOS2058.

Z. Campbell, A. Bray, A. M. Ritz, and A. Groce. Differentially private ANOVA testing. In *ICDIS*, pages 281–285, 2018.

C. L. Canonne, G. Kamath, A. McMillan, A. D. Smith, and J. Ullman. The structure of optimal private tests for simple hypotheses. In *STOC*, pages 310–321, 2019.

C. L. Canonne, G. Kamath, and T. Steinke. The discrete gaussian for differential privacy. In *NeurIPS*, 2020.

K. Chaudhuri and D. J. Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

R. Chetty and J. N. Friedman. A practical method to reduce privacy loss when disclosing statistics based on small samples. *American Economic Review Papers and Proceedings*, 109:414–420, 2019.

R. Chetty, J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter. The opportunity atlas: Mapping the childhood roots of social mobility. Working Paper 25147, National Bureau of Economic Research, October 2018. URL `http://www.nber.org/papers/w25147`.

R. Chetty, N. Hendren, M. R. Jones, and S. R. Porter. Race and Economic Opportunity in the United States: an Intergenerational Perspective*. *The Quarterly Journal of Economics*, 135(2):711–783, 12 2019. URL `https://doi.org/10.1093/qje/qjz042`.

S. Couch, Z. Kazan, K. Shi, A. Bray, and A. Groce. Differentially private nonparametric hypothesis testing. In *CCS*, pages 737–751, 2019.

C. Dwork and J. Lei. Differential privacy and robust statistics. In *STOC*, pages 371–380, 2009.

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503, 2006a.

C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006b.

C. Dwork, A. Smith, T. Steinke, and J. Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84, 2017. doi: 10.1146/annurev-statistics-060116-054123.

E. S. Edgington. *Randomization Tests*, pages 1182–1183. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_56.

G. Evans, G. King, M. Schwenzfeier, and A. Thakurta. Statistically valid inferences from privacy protected data. 2019.

H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Prog. Artif. Intell.*, 2(2-3):113–127, 2014. doi: 10.1007/s13748-013-0040-3.

C. Ferrando, S. Wang, and D. Sheldon. Parametric bootstrap for differentially private confidence intervals. *CoRR*, abs/2006.07749, 2021. URL `https://arxiv.org/abs/2006.07749`.

M. Gaboardi, H. Lim, R. M. Rogers, and S. P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML*, pages 2111–2120, 2016. URL `http://proceedings.mlr.press/v48/rogers16.html`.

A. Gut. *Probability: A Graduate Course*. Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461447078.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. doi: 10.1017/9781139020411.

P. Huber and E. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118210338. URL `https://books.google.com/books?id=j1OhquR_j88C`.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

P. Kairouz, S. Oh, and P. Viswanath. Extremal mechanisms for local differential privacy. *J. Mach. Learn. Res.*, 17:17:1–17:51, 2016. URL `http://jmlr.org/papers/v17/15-135.html`.

R. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York, 2010. ISBN 9780387938394.

J. Lei. Differentially private m-estimators. In *NeurIPS*, pages 361–369, 2011.

I. Mironov. Rényi differential privacy. In *CSF*, pages 263–275, 2017.

R. Rogers and D. Kifer. A new class of private chi-square hypothesis tests. In *AISTATS*, pages 991–1000, 2017.

O. Sheffet. Differentially private ordinary least squares. In *ICML*, pages 3105–3114, 2017.

O. Sheffet. Locally private hypothesis testing. In *ICML*, pages 4612–4621, 2018. URL `http://proceedings.mlr.press/v80/sheffet18a.html`.

J. Stock and M. Watson. *Introduction to Econometrics (3rd edition)*. Addison Wesley Longman, 2011.

A. T. Suresh. Robust hypothesis testing and distribution estimation in hellinger distance. *CoRR*, abs/2011.01848, 2020. URL `https://arxiv.org/abs/2011.01848`.

M. Swanberg, I. Globus-Harris, I. Griffith, A. M. Ritz, A. Groce, and A. Bray. Improved differentially private analysis of variance. *Proc. Priv. Enhancing Technol.*, 2019(3):310–330, 2019. URL `https://doi.org/10.2478/popets-2019-0049`.

L. Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

C. Task and C. Clifton. Differentially private significance testing on paired-sample data. In *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, pages 153–161. SIAM, 2016.

A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Y. Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *UAI*, pages 93–103, 2018.

Y. Wang, J. Lee, and D. Kifer. Differentially private hypothesis testing, revisited. *CoRR*, abs/1511.03376, 2015. URL `http://arxiv.org/abs/1511.03376`.