Differentially Private Linear Regression

Daniel Alabi

Joint work with *Audra McMillan (BU/Northeastern/Apple), Jayshree Sarathy, Adam Smith (BU), and Salil Vadhan*

Main paper: https://arxiv.org/abs/2007.05157

My email: alabid@g.harvard.edu

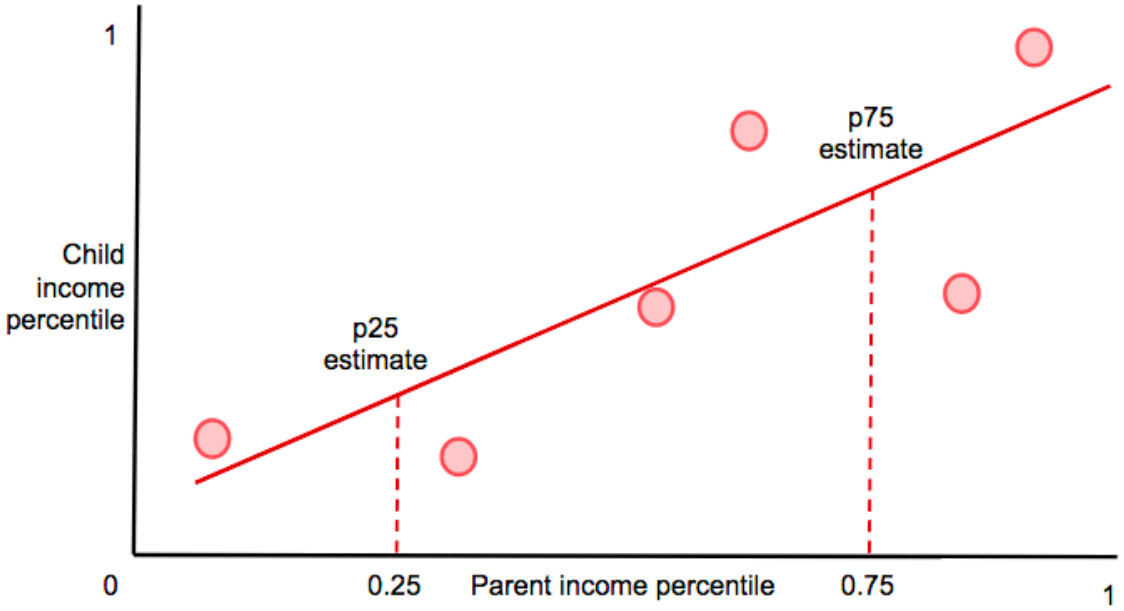
Main Question

Is it possible to design DP linear regression algorithms where the distortion added by the private algorithm is less than the standard error, even for small datasets?

Motivating Application: Opportunity Atlas

Opportunity Insights Application

- Neighborhood-level predictions of social/economic mobility via simple linear regression
- Especially challenging for DP when dataset contains tens to hundreds of datapoints



OI team provide noise infusion algorithm (not formally private) [Chetty-Friedman ‘19] with sufficient accuracy (i.e., error due to privacy less than standard error). We provide DP algorithms for this problem based on robust linear regression estimators (rather than OLS).

Table of Contents

- Definitions
  - Differential Privacy
  - Simple Linear Regression
- Mechanisms
  - DPSuffStats
  - DPGradDescent
  - DPTheilSen
- Experimental Results
  - Opportunity Insights Data
  - Synthetic Datasets

Differential Privacy: Definition (Pure)

[Dwork-McSherry-Nissim-Smith ‘06]

For any algorithm $\mathcal{A}$, it satisfies ϵ-differential privacy if
For all datasets D, D' differing in exactly one row all queries q
Distribution of $\mathcal{A}(D, q)$ is at most ϵ away from $\mathcal{A}(D', q)$

For all sets $T, \epsilon \geq 0$,
$\Pr[\mathcal{A}(D, q) \in T] \leq e^{\epsilon} \Pr[\mathcal{A}(D', q) \in T]$
$e^{\epsilon} \approx 1 + \epsilon$ as $\epsilon \to 0$

Differential Privacy: Definition (Approximate)

[Dwork-McSherry-Nissim-Smith '06]

For any algorithm $\mathcal{A}$, it satisfies (ϵ, δ)-differential privacy if

For all datasets D, D' differing in exactly one row all queries q

For all sets $T, \epsilon \geq 0$,

$\Pr[\mathcal{A}(D, q) \in T] \leq e^{\epsilon} \Pr[\mathcal{A}(D', q) \in T] + \delta$,

$\delta \in [0,\ 1]$. Usually, $\delta \leq n^{-\omega(1)}$.

Differential Privacy: Definition (zCDP/Renyi)

[Bun-Steinke '16, Dwork-Rothblum '16, Mironov '17]

For any algorithm $\mathcal{A}$, it satisfies ρ-zCDP if

For all datasets D, D' differing in exactly one row all queries q

For all $\alpha \in (1, \infty)$,

$$\mathrm{D}_\alpha(\mathcal{A}(D,q) \| \mathcal{A}(D',q)) \leq \rho\alpha\,,$$

where $\mathrm{D}_\alpha(\mathcal{A}(D,q) \| \mathcal{A}(D',q))$ is the α-Renyi divergence between the distribution of $\mathcal{A}(D,q)$ and $\mathcal{A}(D',q)$.

Simple Linear Regression

Assume that:

1) $\forall\, i \in [n],\ y_i = \alpha \cdot x_i + \beta + e_i,$ e_i are error terms

2) $\forall i \in [n],\quad x_i \in \mathbb{R}$

Simple Linear Regression

Assume that:

1) $\forall\, i \in [n],\ y_i = \alpha \cdot x_i + \beta + e_i,$ e_i are error terms

2) $\forall i \in [n],$ $x_i \in \mathbb{R}$

$\hat{\alpha}, \hat{\beta}$ are non-DP estimates of α, β

The goal is to calculate and release DP estimates of:

$\hat{\alpha}, \hat{\beta}$ or $\hat{p}_{25} = 0.25 \cdot \hat{\alpha} + \hat{\beta}, \hat{p}_{75} = 0.75 \cdot \hat{\alpha} + \hat{\beta}.$

OLS (Ordinary Least Squares) Estimator

Assume that:

1) $\forall\, i \in [n],\ y_i = \alpha \cdot x_i + \beta + e_i,$ e_i are error terms

2) $\forall i \in [n],\quad x_i \in \mathbb{R}$

$$X = (x_1, \ldots, x_n)^T, Y = (y_1, \ldots, y_n)^T$$

OLS estimator

$$\hat{\alpha}^{OLS} = ncov(X, Y)/nvar(X)$$

where the sufficient statistics are

- $ncov(X, Y) = (X - \bar{X})^T (Y - \bar{Y})$
- $nvar(X) = (X - \bar{X})^T (X - \bar{X})$

It's the BLUE estimator by the Gauss-Markov theorem.

Theil-Sen Estimator

Theil-Sen estimator (Theil 50, Sen 68):

1) For $i \neq j \in [n], X_i \neq X_j$, compute slopes of pair of points as follows:

$$Z_{ij} = \frac{Y_j - Y_i}{X_j - X_i}.$$

2) Compute median of the Z's.

Breakdown point of ~29%

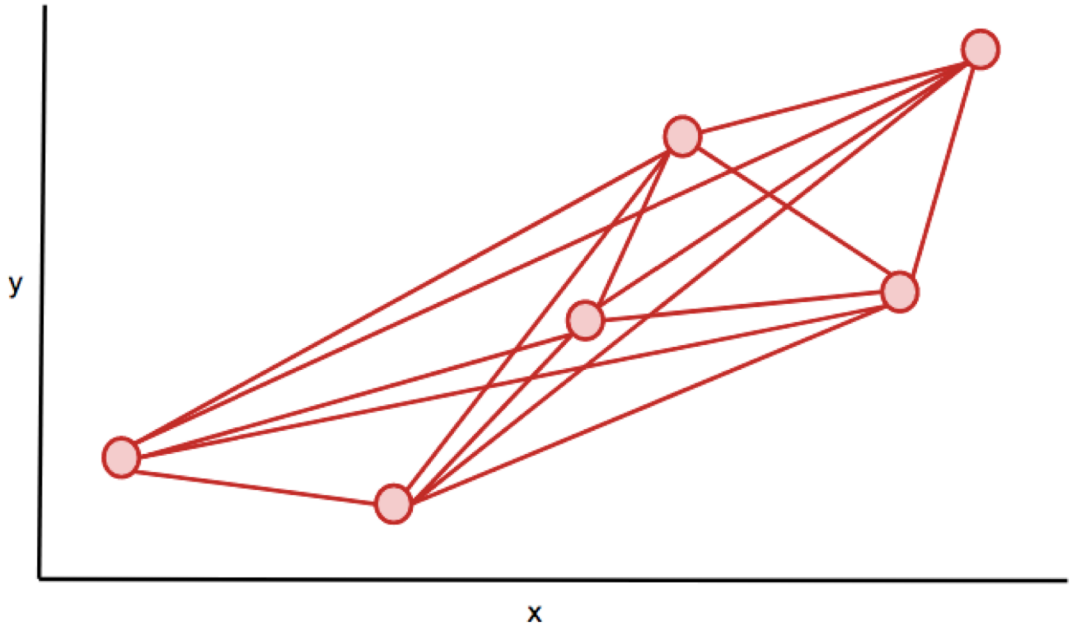


Table of Contents

Robustness and DP Algorithm Design

- Global Sensitivity for query $q: \mathcal{X}^n \to \mathbb{R}^k$ is

$$GS_q = \max_{\{x \sim x' \in \mathcal{X}^n\}} ||q(x) - q(x')||_1$$

- Local Sensitivity for query $q: \mathcal{X}^n \to \mathbb{R}^k$ on dataset $x \in \mathcal{X}^n$ is

$$LS_q = \max_{\{x \sim x' \in \mathcal{X}^n\}} ||q(x) - q(x')||_1$$

Robustness and DP Algorithm Design

- Global Sensitivity for query $q: \mathcal{X}^n \rightarrow \mathbb{R}^k$ is

$$GS_q = \max_{\{x \sim x' \in \mathcal{X}^n\}} ||q(x) - q(x')||_1$$

- Local Sensitivity for query $q: \mathcal{X}^n \rightarrow \mathbb{R}^k$ on dataset $x \in \mathcal{X}^n$ is

$$LS_q = \max_{\{x \sim x' \in \mathcal{X}^n\}} ||q(x) - q(x')||_1$$

For query $q: \mathcal{X}^n \rightarrow \mathbb{R}^k$ for estimating OLS regression parameters, GS_q is infinite. Why?

Robustness and DP Algorithm Design

- Global Sensitivity for query $q: \mathcal{X}^n \to \mathbb{R}^k$ is

$$GS_q = \max_{\{x \sim x' \in \mathcal{X}^n\}} ||q(x) - q(x')||_1$$

- Local Sensitivity for query $q: \mathcal{X}^n \to \mathbb{R}^k$ on dataset $x \in \mathcal{X}^n$ is

$$LS_q = \max_{\{x \sim x' \in \mathcal{X}^n\}} ||q(x) - q(x')||_1$$

For query $q: \mathcal{X}^n \to \mathbb{R}^k$ for estimating OLS regression parameters, GS_q is infinite. Why?

Recall: $\hat{\alpha}^{OLS} = ncov(X, Y)/nvar(X)$

$$ncov(X, Y) = (X - \bar{X})^T (Y - \bar{Y}), nvar(X) = (X - \bar{X})^T (X - \bar{X})$$

Robustness and DP Algorithm Design

To add less noise, we consider DP analogues of robust linear regression estimators (e.g., Theil-Sen, Least Trimmed Squares) rather than OLS.

Robustness and DP Algorithm Design

To add less noise, we consider DP analogues of robust linear regression estimators (e.g., Theil-Sen, Least Trimmed Squares) rather than OLS.

DP Robust methods suited for small datasets

[Dwork-Lei ’09, Couch et al. ‘19]

Robustness and DP Algorithm Design

To add less noise, we consider DP analogues of robust linear regression estimators (e.g., Theil-Sen, Least Trimmed Squares) rather than OLS.

DP Robust methods suited for small datasets

[Dwork-Lei ’09, Couch et al. ‘19]

$$X = (x_1, \ldots, x_n)^T,$$
$$nvar(X) = (X - \bar{X})^T (X - \bar{X})$$

$\epsilon \cdot nvar(X)$ informs whether to choose robust vs. non-robust method

DPSuffStats (closest to OLS estimator)

To make DP, add Laplace noise to the sufficient statistics.

Advantages: as efficient as OLS; can release sufficient statistics (for other tasks); geometric interpretation

DPSuffStats

$$X = (x_1, \ldots, x_n)^T, Y = (y_1, \ldots, y_n)^T$$

OLS estimator

$$\hat{\alpha}^{OLS} = ncov(X, Y)/nvar(X)$$

where the sufficient statistics are

- $ncov(X, Y) = (X - \bar{X})^T (Y - \bar{Y})$
- $nvar(X) = (X - \bar{X})^T (X - \bar{X})$

$$\tilde{\alpha}^{DP} = \begin{cases} \dfrac{ncov(X, Y) + L_1}{nvar(X) + L_2} & \text{if } nvar(X) + L_2 > 0 \\ \perp & otherwise \end{cases}$$

e.g., if data values bounded between 0 and 1, $GS_{ncov} = GS_{nvar} = 1 - \frac{1}{n}$.

$L_1, L_2 \sim \text{Lap}(3(1 - \frac{1}{n})/\epsilon)$

DPSuffStats

$$\forall\, i \in [n], \qquad y_i = \alpha \cdot x_i + \beta + e_i, \qquad e_i \sim \mathcal{N}(0, \sigma_e^2)$$

Ideally, we want the MSE with privacy to approach the MSE without privacy!
Some observations:
1) **Effect of $|\boldsymbol{\alpha}|$:** A smaller $|\boldsymbol{\alpha}|$ leads to a smaller MSE.
2) **Effect of $\boldsymbol{\epsilon}$ (privacy parameter):** A larger $\boldsymbol{\epsilon}$ leads to a smaller MSE.
3) **Effect of $\boldsymbol{n}$ (sample size):** A larger sample size leads to a smaller MSE.
4) **Effect of var(X):** Larger variance of the independent variable leads to a smaller MSE.

MSE without privacy	$\dfrac{\boldsymbol{\sigma_e^2}}{\boldsymbol{n}\mathrm{var}(\boldsymbol{X})}$
MSE with privacy	$\approx \dfrac{\boldsymbol{\sigma_e^2}}{\boldsymbol{n}\mathrm{var}(\boldsymbol{X})} + \dfrac{\lvert\boldsymbol{\alpha}\rvert}{\left(\boldsymbol{\epsilon n}\mathrm{var}(\boldsymbol{X})\right)^2}$

DPGradDescent

Convex optimization problem that defines OLS:

Minimize $||Y - (\alpha \cdot X + \beta)||^2$

Solve using gradient descent.

To make DP, add noise to the gradient computation process.

Advantages: inherits most benefits of non-private gradient descent (e.g., parallelizability, fine-tuning of optimization steps)

DPGradDescent

Convex optimization problem that defines OLS:

Minimize $||Y - (\alpha \cdot X + \beta)||^2$

Solve using gradient descent.

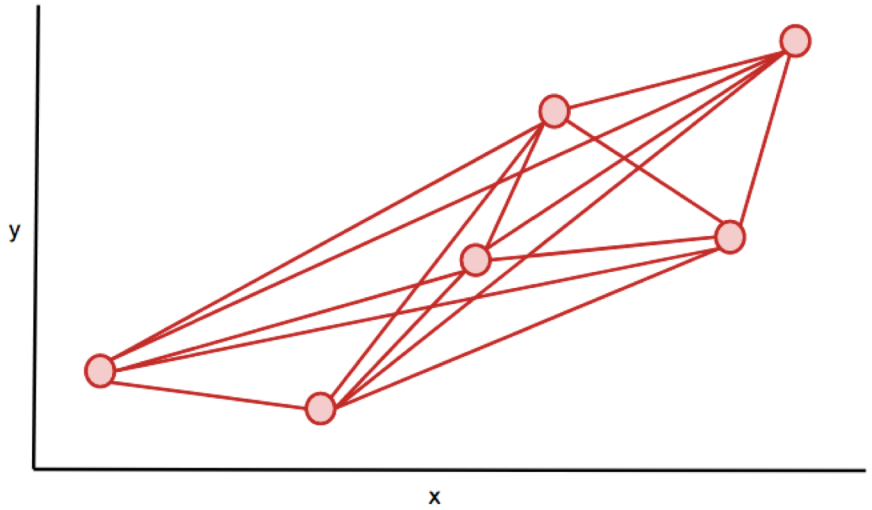
To make DP, add noise to the gradient computation process.

We use Laplace (fatter tails), Gaussian noise and have to clip gradients.

DPTheilSen

Theil-Sen estimator (Theil 50, Sen 68):
1) For $i \neq j \in [n], X_i \neq X_j$, compute slopes of pair of points as follows:
$Z_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$.
2) Compute median of the Z's.

To make DP, make median computation DP.



Advantages: performs best on smallest datasets (e.g., tens or hundreds); as DP median computation gets better (i.e., more accurate, faster), DPTheilSen gets better.

DPTheilSen

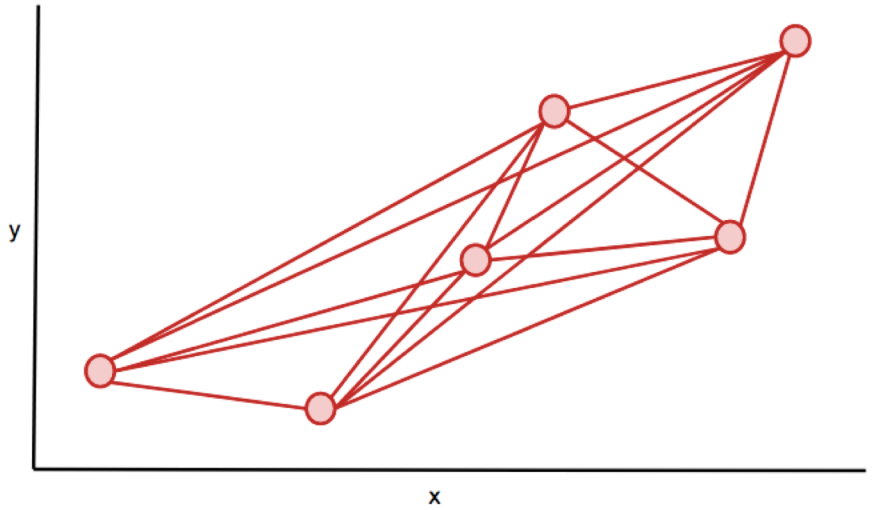
Theil-Sen estimator (Theil 50, Sen 68):

1) For $i \neq j \in [n], X_i \neq X_j$, compute slopes of pair of points as follows:

$Z_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$.

2) Compute median of the Z's.

To make DP, make median computation DP.



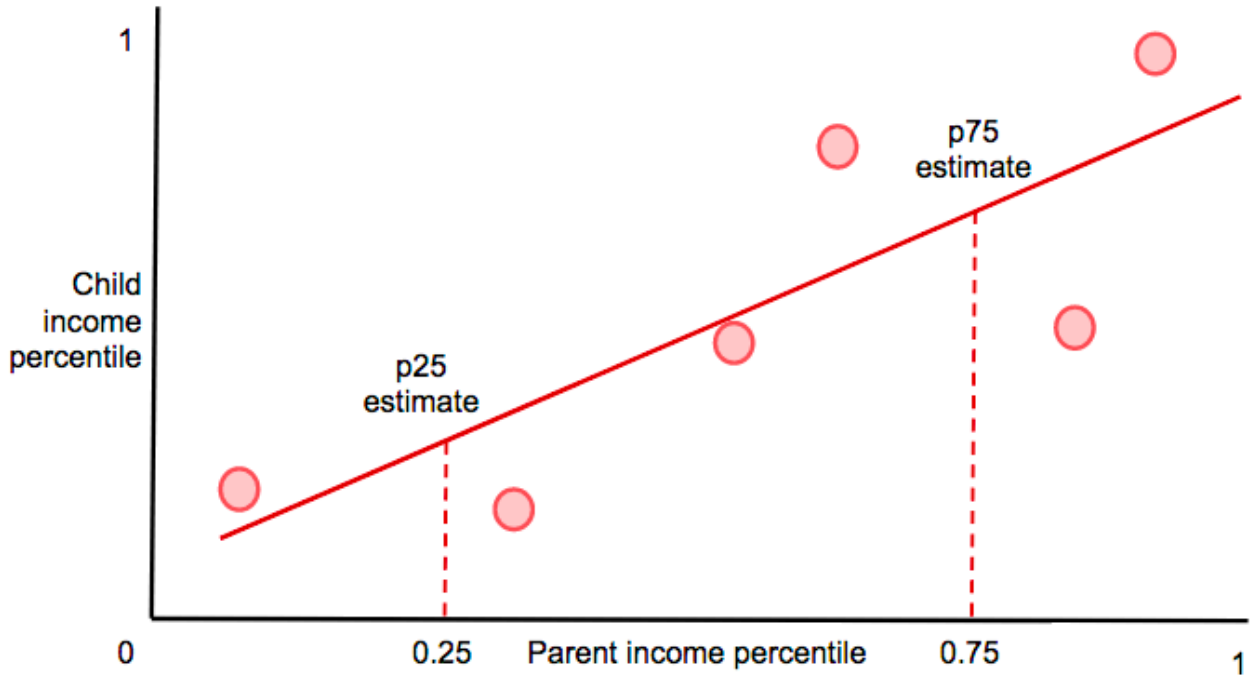
Main DP Median computation:

1) DPSSTheilSen: Smooth sensitivity upper bound on local sensitivity of median. [Nissim et al. '07]
2) DPExpTheilSen: Exponential mechanism for median. [McSherry-Talwar '07]

Table of Contents

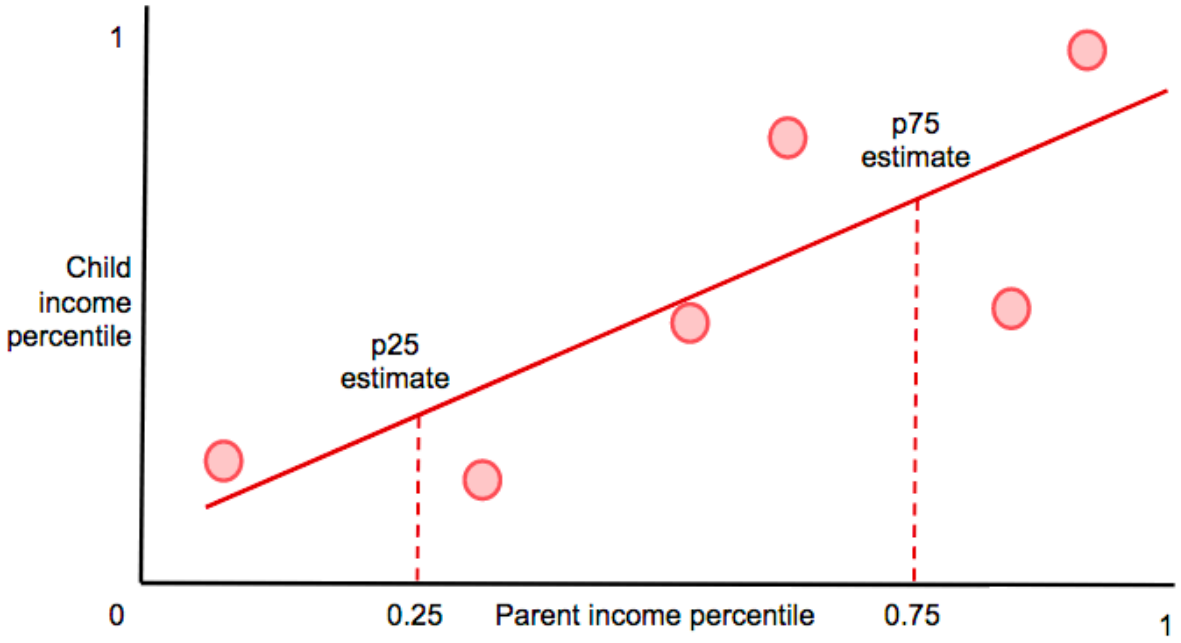
Opportunity Insights Application

- Neighborhood-level predictions of social mobility via simple linear regression



Opportunity Insights Application

- Neighborhood-level predictions of social mobility via simple linear regression



They provide noise infusion algorithm (not formally private) [Chetty-Friedman ‘19] with sufficient accuracy (i.e., error due to privacy less than standard error).

Error Metrics

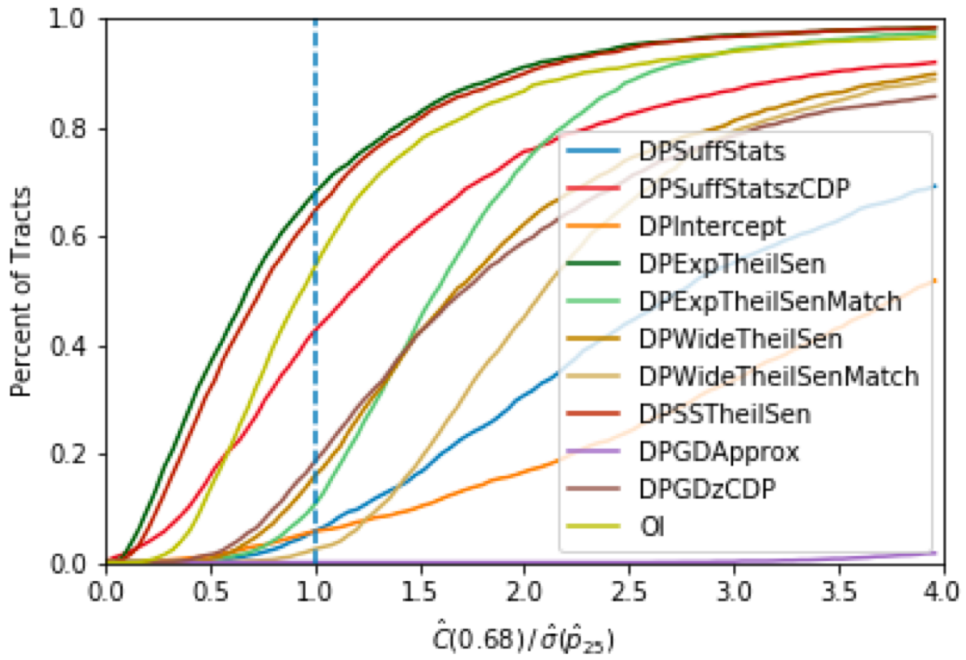
Recall that the goal is to calculate and release DP estimates of:

$\hat{\alpha}, \hat{\beta}$ or $\hat{p}_{25} = 0.25 \cdot \hat{\alpha} + \hat{\beta}, \hat{p}_{75} = 0.75 \cdot \hat{\alpha} + \hat{\beta}$.

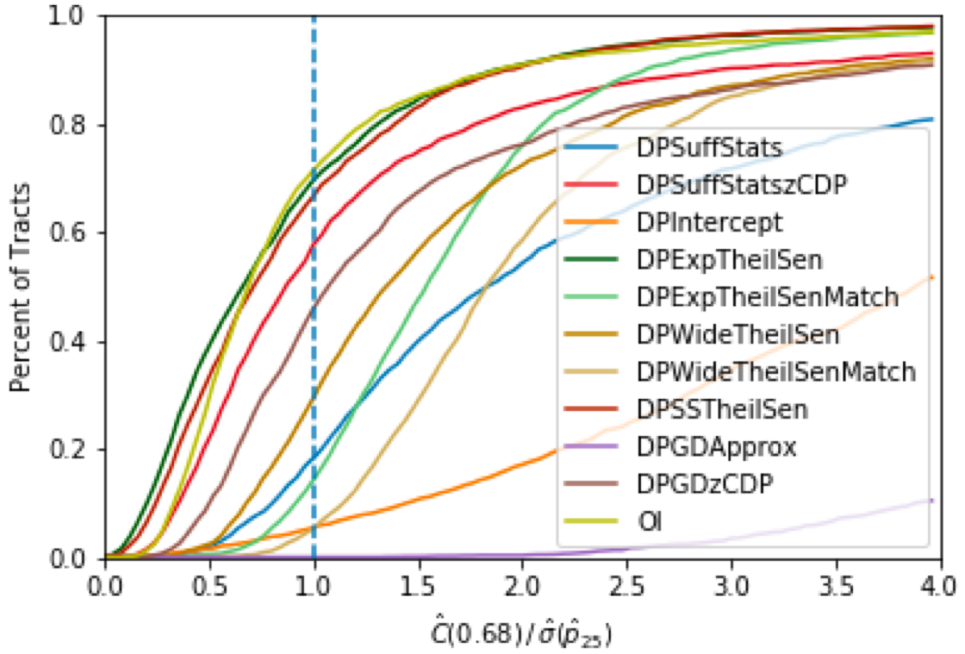
We use tildes to denote private estimates and hats for non-private estimates.

- $C(q) = \min\{c : \mathbb{P}(|\tilde{p}_{25} - \hat{p}_{25}| \leq c) \geq q\}$ for any $q \in [0, 1]$
- $\hat{C}(q) = \min\{c : \geq q\ fraction\ of\ trials\ have\ error \leq c\}$ for any $q \in [0, 1]$
- $\sigma(p) =$ standard deviation of estimator p under a noise model (i.e., error terms gaussian)
- $\hat{\sigma}(p) =$ standard error of estimator p = empirical estimate of $\sigma(p)$

Opportunity Insights (OI) Application (IL, NC)



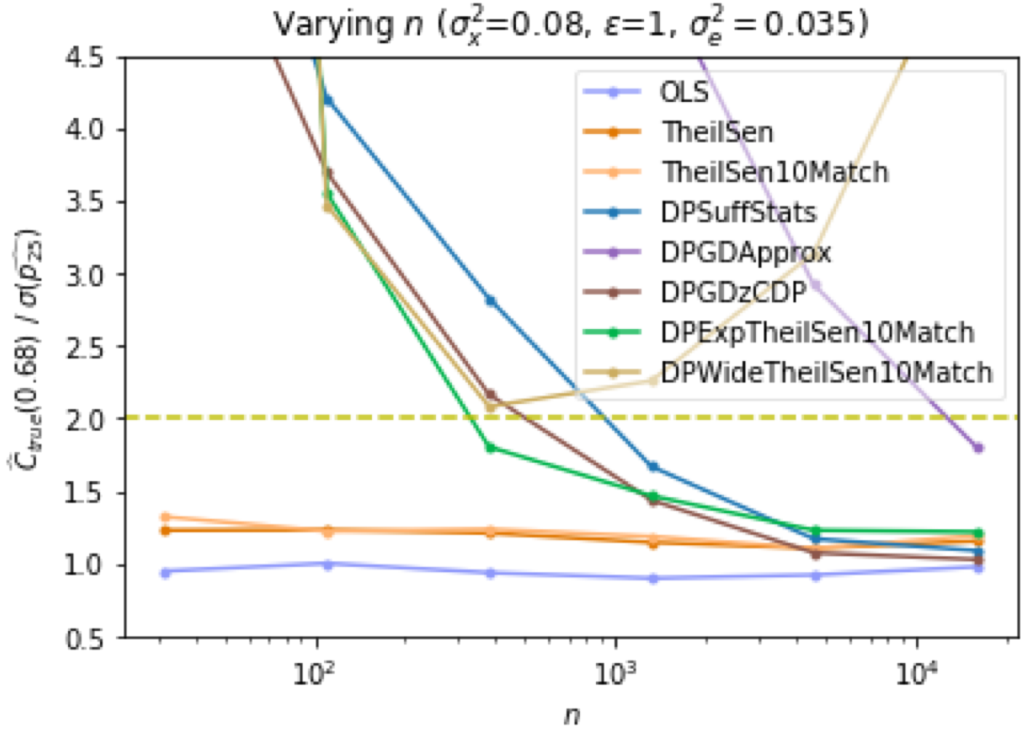
Illinois

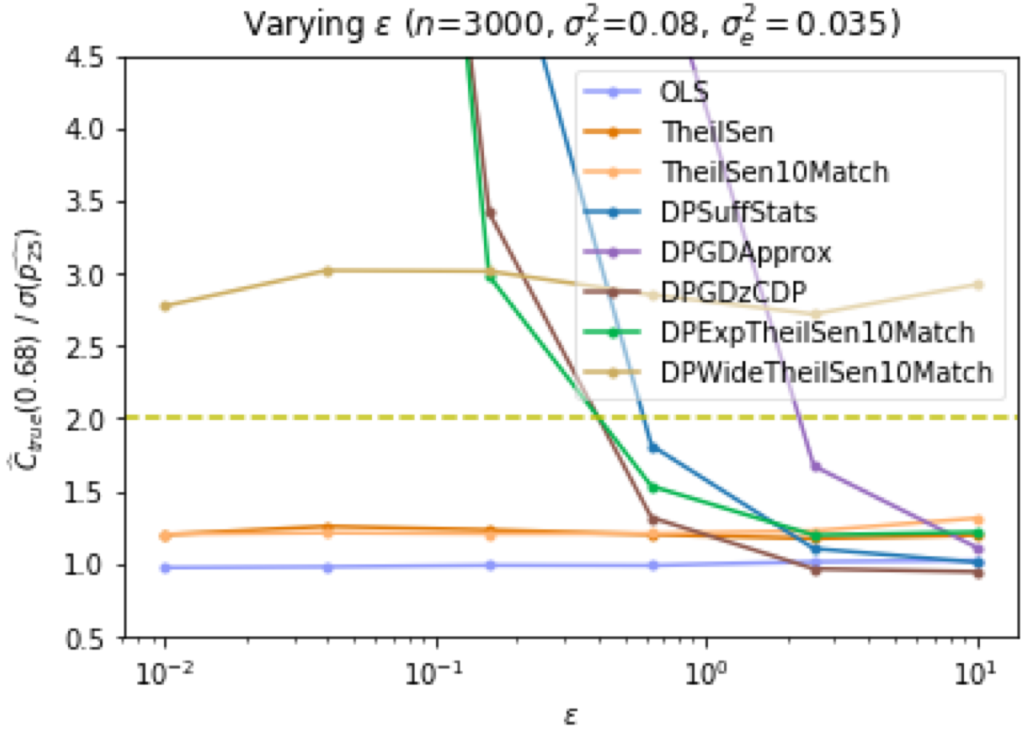


North Carolina

Main Takeaway: DPExpTheilSen outperforms or matches the OI method

Synthetic Datasets





Main Takeaway: Private robust methods (e.g., DPExpTheilSen) approach non-private robust methods; Private non-robust methods (e.g., DPSuffStats) approach non-private non-robust methods; there's a cross-over between private robust/non-robust methods

Conclusion

- It is possible to design DP simple linear regression algorithms where the distortion added by the private algorithm is less than the standard error, even for small datasets.

- DP OLS-type estimators work great when $\epsilon \cdot nvar(X)$ large.

- Otherwise, DP robust linear regression estimators perform better.

Future Work. Any Questions?

	[DL09]	[ZZX+12]	[DJW13]	[BST14]	[She15]	[She17]
Uses Bayesian Approach?	No	No	No	No	No	No
Point Estimates?	Yes	Yes	Yes	Yes	Yes	Yes
Uncertainty Estimates?	No	No	No	No	No	Yes
Multiple Linear Regression?	Yes	Yes	Yes	Yes	Yes	Yes
Small Dataset (e.g., ≤ 500)	Yes	No	No	No	No	No
Ridge/OLS/Robust Estimator	Robust	OLS	OLS	OLS	Ridge	OLS/Ridge
Distributed?	No	No	Yes	No	No	No

	[STU17]	[Wan18]	[CKS+19]	[BS19]	[CF19]	[AMS+20]
Uses Bayesian Approach?	No	No	No	Yes	No	No
Point Estimates?	Yes	Yes	Yes	Yes	Yes	Yes
Uncertainty Estimates?	No	No	Yes	No	No	No
Multiple Linear Regression?	Yes	Yes	Yes	Yes	No	No
Small Dataset (e.g., ≤ 500)	No	No	No	No	Yes	Yes
Ridge/OLS/Robust Estimator	OLS	Ridge	OLS/Robust	OLS	OLS	OLS/Robust
Distributed?	Yes	No	No	No	No	No