Robustness of OLS to sample removals: theoretical analysis and implications

Boaz Nadler

Weizmann Institute of Science, Israel

Joint work with
Michael Feldman and Eyar Azar

Vienna, Sep. 2025

Outline

- Trust in a learned model; specifically its robustness to removal of few samples

Outline

- Trust in a learned model; specifically its robustness to removal of few samples

- Most influential subset selection problem / robustness auditing

Outline

- Trust in a learned model; specifically its robustness to removal of few samples

- Most influential subset selection problem / robustness auditing

- A real data example

Outline

- Trust in a learned model; specifically its robustness to removal of few samples

- Most influential subset selection problem / robustness auditing

- A real data example

- Theoretical analysis of robustness auditing for ordinary least squares

Outline

- Trust in a learned model; specifically its robustness to removal of few samples

- Most influential subset selection problem / robustness auditing

- A real data example

- Theoretical analysis of robustness auditing for ordinary least squares

- Revisit the real data example

Outline

- Trust in a learned model; specifically its robustness to removal of few samples

- Most influential subset selection problem / robustness auditing

- A real data example

- Theoretical analysis of robustness auditing for ordinary least squares

- Revisit the real data example

- Insights and implications

Statistical Learning Paradigm

Standard workflow of supervised learning:

Statistical Learning Paradigm

Standard workflow of supervised learning:

Input:

training set of n samples $(\mathbf{x}_i, y_i), \quad i = 1, \ldots, n$

Statistical Learning Paradigm

Standard workflow of supervised learning:

Input:

training set of n samples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$

Goal:

construct a predictor for the response y for new $\mathbf{x}$'s whose responses y are not observed

Statistical Learning Paradigm

Standard workflow of supervised learning:

Input:

training set of n samples $(\mathbf{x}_i, y_i), \quad i = 1, \ldots, n$

Goal:

construct a predictor for the response y for new $\mathbf{x}$'s whose responses y are not observed

Focus in this talk: regression setting

Statistical Learning Paradigm

Standard workflow of supervised learning:

Input:

training set of n samples $(\mathbf{x}_i, y_i)$, $\;i = 1, \ldots, n$

Goal:

construct a predictor for the response y for new $\mathbf{x}$'s whose responses y are not observed

Focus in this talk: regression setting

$$\mathbf{x} \in \mathbb{R}^p \qquad y \in \mathbb{R}$$

Statistical Learning Paradigm

Standard workflow of supervised learning:

Input:

training set of n samples $(\mathbf{x}_i, y_i)$, $\;i = 1, \ldots, n$

Goal:

construct a predictor for the response y for new $\mathbf{x}$'s whose responses y are not observed

Focus in this talk: regression setting

$$\mathbf{x} \in \mathbb{R}^p \qquad y \in \mathbb{R}$$

Assume:

p = number of features $<$ n = number of samples

X - $n \times p$ matrix of all sample is of full rank p

Constructing a Regression Model

Predictor f often found by choosing a loss function ℓ, and minimizing empirical risk over some class of functions $\mathcal{F}$,

Constructing a Regression Model

Predictor f often found by choosing a loss function ℓ, and minimizing empirical risk over some class of functions $\mathcal{F}$,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i))$$

Constructing a Regression Model

Predictor f often found by choosing a loss function ℓ, and minimizing empirical risk over some class of functions $\mathcal{F}$,

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i))$$

Example: Squared loss combined with class $\mathcal{F}$ of linear functions gives *ordinary least squares*,

$$\hat{f}(\mathbf{x}) = \widehat{\boldsymbol{\beta}}^\top \mathbf{x}$$

Constructing a Regression Model

Predictor f often found by choosing a loss function ℓ, and minimizing empirical risk over some class of functions $\mathcal{F}$,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i))$$

Example: Squared loss combined with class $\mathcal{F}$ of linear functions gives *ordinary least squares*,

$$\hat{f}(\mathbf{x}) = \widehat{\boldsymbol{\beta}}^\top \mathbf{x}$$

where

$$\widehat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \boldsymbol{y}$$

Trust in Learned Model

Can we trust learned model / its predictions?

Trust in Learned Model

Can we trust learned model / its predictions?

Accuracy of predictions: cross validation, conformal prediction

Trust in Learned Model

Can we trust learned model / its predictions?

Accuracy of predictions: cross validation, conformal prediction

Accuracy of estimated model parameters: asymptotic confidence intervals, bootstrap.

Trust in Learned Model

Can we trust learned model / its predictions?

Accuracy of predictions: cross validation, conformal prediction

Accuracy of estimated model parameters: asymptotic confidence intervals, bootstrap.

Robustness: detection of outliers, or very influential (individual) samples.

Trust in Learned Model

Can we trust learned model / its predictions?

Accuracy of predictions: cross validation, conformal prediction

Accuracy of estimated model parameters: asymptotic confidence intervals, bootstrap.

Robustness: detection of outliers, or very influential (individual) samples.

In the statistics literature: long history of works under umbrella of

Robustness Diagnostics

Robustness Auditing

In past few years, several researchers noted that in various datasets, removing a *small* number $k \ll n$ of (specifically chosen) training samples leads to *large* changes in the learned model,

Robustness Auditing

In past few years, several researchers noted that in various datasets, removing a *small* number $k \ll n$ of (specifically chosen) training samples leads to *large* changes in the learned model,

very different predictions and/or estimated coefficients

Robustness Auditing

In past few years, several researchers noted that in various datasets, removing a *small* number $k \ll n$ of (specifically chosen) training samples leads to *large* changes in the learned model,

very different predictions and/or estimated coefficients

Example: For OLS, let $\mathcal{S} \subset [n]$ be remaining set of samples after removal of k specific samples, $|\mathcal{S}| = n - k$,

$$\widehat{\beta}_{\mathcal{S}} = \left(X_{\mathcal{S}}^{\top} X_{\mathcal{S}}\right)^{-1} X_{\mathcal{S}}^{\top} \boldsymbol{y}_{\mathcal{S}}.$$

Robustness Auditing

In past few years, several researchers noted that in various datasets, removing a *small* number $k \ll n$ of (specifically chosen) training samples leads to *large* changes in the learned model,

very different predictions and/or estimated coefficients

Example: For OLS, let $\mathcal{S} \subset [n]$ be remaining set of samples after removal of k specific samples, $|\mathcal{S}| = n - k$,

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}} = \left(X_{\mathcal{S}}^{\top} X_{\mathcal{S}}\right)^{-1} X_{\mathcal{S}}^{\top} \boldsymbol{y}_{\mathcal{S}}.$$

If for a specific coefficient $j \in [p]$, with $k \ll n$ samples removed,

$$\widehat{\boldsymbol{\beta}}_j > 0 \quad \text{but} \quad (\widehat{\boldsymbol{\beta}}_{\mathcal{S}})_j < 0$$

our trust in the model may be questionable

Robustness Auditing

In recent years several authors emphasized the need to assess how subsets of training samples collectively affect a learned model,

Robustness Auditing

In recent years several authors emphasized the need to assess how subsets of training samples collectively affect a learned model,

- Koh et al, On the accuracy of influence functions for measuring group effects, NeurIPS 19'
- Basu et al, On second-order group influence functions for black-box predictions, ICML 20'
- Hu et al, Most influential subset selection: Challenges, promises, and beyond, NeurIPS 24'

Robustness Auditing

In recent years several authors emphasized the need to assess how subsets of training samples collectively affect a learned model,

- Koh et al, On the accuracy of influence functions for measuring group effects, NeurIPS 19'
- Basu et al, On second-order group influence functions for black-box predictions, ICML 20'
- Hu et al, Most influential subset selection: Challenges, promises, and beyond, NeurIPS 24'

Broderick et al, 2020, considered a more stringent (worst-case) form of robustness, called robustness auditing:

Robustness Auditing

In recent years several authors emphasized the need to assess how subsets of training samples collectively affect a learned model,

- Koh et al, On the accuracy of influence functions for measuring group effects, NeurIPS 19'
- Basu et al, On second-order group influence functions for black-box predictions, ICML 20'
- Hu et al, Most influential subset selection: Challenges, promises, and beyond, NeurIPS 24'

Broderick et al, 2020, considered a more stringent (worst-case) form of robustness, called robustness auditing:

estimated parameters / predictions need to be robust to removal of *any* subset of k samples

Most influential subset selection problem

[Rubinstein and Hopkins 25']

Definition (OLS): Robustness of $\widehat{\beta}$ to k sample removals in a fixed direction $\boldsymbol{v}$ is measured by

$$\Delta_k(\boldsymbol{v}) = \max_{\mathcal{S}\subseteq[n],|\mathcal{S}|=n-k} \langle \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}, \boldsymbol{v} \rangle$$

Most influential subset selection problem

[Rubinstein and Hopkins 25']

Definition (OLS): Robustness of $\widehat{\boldsymbol{\beta}}$ to k sample removals in a fixed direction $\boldsymbol{v}$ is measured by

$$\Delta_k(\boldsymbol{v}) = \max_{\mathcal{S} \subseteq [n], |\mathcal{S}| = n-k} \langle \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}, \boldsymbol{v} \rangle$$

In particular $\Delta_k(\boldsymbol{e}_j)$ measures sensitivity of j-th coefficient to removal of k samples.

Most influential subset selection problem

[Rubinstein and Hopkins 25']

Definition (OLS): Robustness of $\widehat{\beta}$ to k sample removals in a fixed direction $\boldsymbol{v}$ is measured by

$$\Delta_k(\boldsymbol{v}) = \max_{\mathcal{S}\subseteq[n],|\mathcal{S}|=n-k} \langle \widehat{\beta} - \widehat{\beta}_{\mathcal{S}}, \boldsymbol{v} \rangle$$

In particular $\Delta_k(\boldsymbol{e}_j)$ measures sensitivity of j-th coefficient to removal of k samples.

Key Questions:

- Practical: For a given dataset and a given k, how large can $\Delta_k(\boldsymbol{v})$ be ?

Most influential subset selection problem

[Rubinstein and Hopkins 25']

Definition (OLS): Robustness of $\widehat{\beta}$ to k sample removals in a fixed direction $\boldsymbol{v}$ is measured by

$$\Delta_k(\boldsymbol{v}) = \max_{\mathcal{S}\subseteq[n],|\mathcal{S}|=n-k} \langle \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}, \boldsymbol{v} \rangle$$

In particular $\Delta_k(\boldsymbol{e}_j)$ measures sensitivity of j-th coefficient to removal of k samples.

Key Questions:

- Practical: For a given dataset and a given k, how large can $\Delta_k(\boldsymbol{v})$ be ?

- Theoretical: How large should we expect $\Delta_k(\boldsymbol{v})$ to be under reasonable assumptions about the data?

Robustness Auditing

- Exact computation of $\Delta_k(\boldsymbol{v})$: $\binom{n}{k}$ subsets is computationallly intractable

Robustness Auditing

- Exact computation of $\Delta_k(\boldsymbol{v})$: $\binom{n}{k}$ subsets is computationallly intractable

approximations as well as upper/lower bounds for $\Delta_k(\boldsymbol{v})$

- Exact computation of $\Delta_k(\boldsymbol{v})$: $\binom{n}{k}$ subsets is computationallly intractable

approximations as well as upper/lower bounds for $\Delta_k(\boldsymbol{v})$

- Broderick, T., Giordano, R., and Meager, `AMIP`, 20'
- Kuschnig, Zens and Cuaresma, 21'.
- Moitra and Rohatgi, 23'
- Rubinstein and Hopkins, `ACRE`, 25'

Robustness Auditing

- Exact computation of $\Delta_k(\boldsymbol{v})$: $\binom{n}{k}$ subsets is computationallly intractable

approximations as well as upper/lower bounds for $\Delta_k(\boldsymbol{v})$

- Broderick, T., Giordano, R., and Meager, `AMIP`, 20'
- Kuschnig, Zens and Cuaresma, 21'.
- Moitra and Rohatgi, 23'
- Rubinstein and Hopkins, `ACRE`, 25'

Several of these works show that on various datasets, OLS is *not* robust to removal of even just a handful of samples.

Real Data Example: Cash transfers datasets

[Angelucci and De Giorgi 2009']

Econometric study on Mexico's progresa aid program

Real Data Example: Cash transfers datasets

[Angelucci and De Giorgi 2009']

Econometric study on Mexico's progresa aid program

506 rural villages: 320 participated in program, 186 control group.

Real Data Example: Cash transfers datasets

[Angelucci and De Giorgi 2009']

Econometric study on Mexico's progresa aid program

506 rural villages: 320 participated in program, 186 control group.

Poor households in participating villages received financial support

Real Data Example: Cash transfers datasets

[Angelucci and De Giorgi 2009']

Econometric study on Mexico's progresa aid program

506 rural villages: 320 participated in program, 186 control group.

Poor households in participating villages received financial support

Effect of program estimated by a linear regression

$$y = \beta_0 + \beta_1 x_1 + \sum_{j=2}^{17} \beta_j x_j,$$

Real Data Example: Cash transfers datasets

[Angelucci and De Giorgi 2009']

Econometric study on Mexico's progresa aid program

506 rural villages: 320 participated in program, 186 control group.

Poor households in participating villages received financial support

Effect of program estimated by a linear regression

$$y = \beta_0 + \beta_1 x_1 + \sum_{j=2}^{17} \beta_j x_j,$$

response y - total household consumption in pesos,

Real Data Example: Cash transfers datasets

[Angelucci and De Giorgi 2009']

Econometric study on Mexico's progresa aid program

506 rural villages: 320 participated in program, 186 control group.

Poor households in participating villages received financial support

Effect of program estimated by a linear regression

$$y = \beta_0 + \beta_1 x_1 + \sum_{j=2}^{17} \beta_j x_j,$$

response y - total household consumption in pesos,
x_1 - binary treatment variable (1 if village participated in Progresa),
$x_2, \ldots, x_{17}$ additional covariates of each household.

Real Data Example: Cash transfers datasets

[Angelucci and De Giorgi 2009']

Econometric study on Mexico's progresa aid program

506 rural villages: 320 participated in program, 186 control group.

Poor households in participating villages received financial support

Effect of program estimated by a linear regression

$$y = \beta_0 + \beta_1 x_1 + \sum_{j=2}^{17} \beta_j x_j,$$

response y - total household consumption in pesos,
x_1 - binary treatment variable (1 if village participated in Progresa),
$x_2, \ldots, x_{17}$ additional covariates of each household.
In non-poor households, β_1 captures (indirect) effect of the Progresa aid program.

Cash Transfers Dataset

Period	Poor	n	$\widehat{\beta}_1$	AMIP
8	Y	10781	16.53	225
8	N	4543	-5.53	5
9	Y	9489	28.65	321
9	N	3769	23.19	21
10	Y	10368	32.52	570
10	N	4191	21.12	26

6 rows: three time periods $\times$ two groups (poor/non poor)

Cash Transfers Dataset

Period	Poor	n	$\widehat{\beta}_1$	AMIP
8	Y	10781	16.53	225
8	N	4543	-5.53	5
9	Y	9489	28.65	321
9	N	3769	23.19	21
10	Y	10368	32.52	570
10	N	4191	21.12	26

6 rows: three time periods $\times$ two groups (poor/non poor)

Column 5: size of smallest subset found by AMIP whose removal changes the sign of $\widehat{\beta}_1$

Some Questions

How come only $k = 5 - 30$ samples out of $n > 4000$ can change the sign of an OLS coefficient? (even when it is considered as statistically significant by standard tests)

Some Questions

How come only $k = 5 - 30$ samples out of $n > 4000$ can change the sign of an OLS coefficient? (even when it is considered as statistically significant by standard tests)

This talk:

Some Questions

How come only $k = 5 - 30$ samples out of $n > 4000$ can change the sign of an OLS coefficient? (even when it is considered as statistically significant by standard tests)

This talk:

Theoretical analysis: robustness auditing of OLS

Some Questions

How come only $k = 5 - 30$ samples out of $n > 4000$ can change the sign of an OLS coefficient? (even when it is considered as statistically significant by standard tests)

This talk:

Theoretical analysis: robustness auditing of OLS

- Derive conditions (and understanding) when would OLS be provably robust to k-sample removals

Some Questions

How come only $k = 5 - 30$ samples out of $n > 4000$ can change the sign of an OLS coefficient? (even when it is considered as statistically significant by standard tests)

This talk:

Theoretical analysis: robustness auditing of OLS

- Derive conditions (and understanding) when would OLS be provably robust to k-sample removals

- conversely, when would OLS be provably *non-robust* to k sample removals

Some Questions

How come only $k = 5 - 30$ samples out of $n > 4000$ can change the sign of an OLS coefficient? (even when it is considered as statistically significant by standard tests)

This talk:

Theoretical analysis: robustness auditing of OLS

- Derive conditions (and understanding) when would OLS be provably robust to k-sample removals

- conversely, when would OLS be provably *non-robust* to k sample removals

- Revisit the cash transfer dataset and identify potential causes

Some Questions

How come only $k = 5 - 30$ samples out of $n > 4000$ can change the sign of an OLS coefficient? (even when it is considered as statistically significant by standard tests)

This talk:

Theoretical analysis: robustness auditing of OLS

- Derive conditions (and understanding) when would OLS be provably robust to k-sample removals

- conversely, when would OLS be provably *non-robust* to k sample removals

- Revisit the cash transfer dataset and identify potential causes

- Implications for practitioners

Robustness Auditing of OLS

Question:

When would OLS be robust to removal of k samples?

Question:

When would OLS be robust to removal of k samples?

Assume samples $(\mathbf{x}_i, y_i)$ are i.i.d. from joint distribution $P(\mathbf{x}, y)$,

Robustness Auditing of OLS

Question:

When would OLS be robust to removal of k samples?

Assume samples $(\mathbf{x}_i, y_i)$ are i.i.d. from joint distribution $P(\mathbf{x}, y)$,

Analyze two models for $P(\mathbf{x}, y)$:

Robustness Auditing of OLS

Question:

When would OLS be robust to removal of k samples?

Assume samples $(\mathbf{x}_i, y_i)$ are i.i.d. from joint distribution $P(\mathbf{x}, y)$,

Analyze two models for $P(\mathbf{x}, y)$:

- Model 1: General P with mild regularity conditions

Robustness Auditing of OLS

Question:

When would OLS be robust to removal of k samples?

Assume samples $(\mathbf{x}_i, y_i)$ are i.i.d. from joint distribution $P(\mathbf{x}, y)$,

Analyze two models for $P(\mathbf{x}, y)$:

- Model 1: General P with mild regularity conditions
- Model 2: Gaussian Linear Model.

OLS Robustness

Model 1 - General

$P(\mathbf{x}, y)$ - probability distribution over $\mathbb{R}^{p+1}$ such that

1. $\mathbf{x}$ is sub-Gaussian, zero mean, and has positive-definite covariance $\Sigma = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$
2. y is sub-Gaussian

OLS Robustness

Model 2 - Gaussian Linear

$y = \boldsymbol{\beta}^\top \cdot \mathbf{x} + \varepsilon$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is deterministic and

1. $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$, $\Sigma \succ 0$ is positive definite
2. ε is sub-Gaussian, mean zero, and independent of $\mathbf{x}$

Preliminaries

Optimal prediction under squared loss is *conditional mean*

Preliminaries

Optimal prediction under squared loss is *conditional mean*

$$\mathbb{E}[y \mid \mathbf{x}]$$

Preliminaries

Optimal prediction under squared loss is *conditional mean*

$$\mathbb{E}[y \mid \mathbf{x}]$$

Optimal OLS solution is

$$\beta^{\mathsf{OLS}} = \Sigma^{-1} \cdot \mathbb{E}[\mathbf{x}y]$$

Preliminaries

Optimal prediction under squared loss is *conditional mean*

$$\mathbb{E}[y \mid \mathbf{x}]$$

Optimal OLS solution is

$$\beta^{\text{OLS}} = \Sigma^{-1} \cdot \mathbb{E}[\mathbf{x}y]$$

The general model 1 is *mis-specified*, since $\mathbb{E}[y \mid \mathbf{x}]$ may be a non-linear function of $\mathbf{x}$

Preliminaries

Optimal prediction under squared loss is *conditional mean*

$$\mathbb{E}[y \mid \mathbf{x}]$$

Optimal OLS solution is

$$\boldsymbol{\beta}^{\mathsf{OLS}} = \Sigma^{-1} \cdot \mathbb{E}[\mathbf{x}y]$$

The general model 1 is *mis-specified*, since $\mathbb{E}[y \mid \mathbf{x}]$ may be a non-linear function of $\mathbf{x}$

Under Gaussian-Linear Model 2, $\boldsymbol{\beta}^{\mathsf{OLS}} = \boldsymbol{\beta}$

Theoretical Results under Model 1

Sub-Gaussian norm of y:

$$\|y\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}[\exp(|y|^2/t^2)] \le e\right\}.$$

Theoretical Results under Model 1

Sub-Gaussian norm of y:

$$\|y\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}[\exp(|y|^2/t^2)] \le e\right\}.$$

Sub-Gaussian norm of vector x:

$$\|\mathbf{x}\|_{\psi_2} = \sup_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \|\boldsymbol{v}^\top \mathbf{x}\|_{\psi_2},$$

Non-asymptotic bound

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ i.i.d. from Model 1.

Non-asymptotic bound

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ i.i.d. from Model 1.

Theorem

Assume $k \leq n/2$, $p \leq n - k$, and define

$$\eta = \|\Sigma^{-1/2}\| \left(1 + \|\Sigma^{-1/2}\mathbf{x}\|_{\psi_2}^4\right) \|y\|_{\psi_2}^2.$$

There exist constants $C, c > 0$ such that with probability at least $1 - 7(n/k)^{-ck}$,

$$\max_{\mathcal{S} \subseteq [n] : |\mathcal{S}| \geq n-k} \|\widehat{\beta}_{\mathcal{S}} - \widehat{\beta}\| \leq C\eta \sqrt{\frac{k \log(en/k)}{n-k}}$$

for all p, n, k satisfying

$$C \|\Sigma^{-1/2}\mathbf{x}\|_{\psi_2}^2 \sqrt{\frac{p}{n-k}} \leq 1.$$

OLS Robustness Under Model 1

Asymptotics as $n \to \infty$, $k = k_n$ and $p = p_n$ may tend to infinity.

OLS Robustness Under Model 1

Asymptotics as $n \to \infty$, $k = k_n$ and $p = p_n$ may tend to infinity.

Assume $\|\Sigma^{-1/2}\|$, $\|\mathbf{x}\|_{\psi_2}$, and $\|y\|_{\psi_2}$ remain bounded as $n \to \infty$.

OLS Robustness Under Model 1

Asymptotics as $n \to \infty$, $k = k_n$ and $p = p_n$ may tend to infinity.

Assume $\|\Sigma^{-1/2}\|$, $\|\mathbf{x}\|_{\psi_2}$, and $\|y\|_{\psi_2}$ remain bounded as $n \to \infty$.

Theorem

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ *i.i.d. from Model 1.*
Assume that as $n \to \infty$, $\limsup \|\Sigma^{-1/2}\mathbf{x}\|_{\psi_2}^2 \sqrt{p/n} < c$, *where* $c > 0$ *is an absolute constant. Then, if* $k/n \to 0$,

$$\max_{\mathcal{S}\subseteq[n], |\mathcal{S}|\geq n-k} \|\widehat{\beta}_{\mathcal{S}} - \widehat{\beta}\| \xrightarrow{p} 0$$

OLS Robustness Under Model 1

In simple words: for "well-behaved" data, if $k/n \to 0$, OLS is robust to k sample removals, regardless of model mis-specification, and regardless of data dimension.

OLS Robustness Under Model 1

In simple words: for “well-behaved” data, if $k/n \to 0$, OLS is robust to k sample removals, regardless of model mis-specification, and regardless of data dimension.

In fact, theorem allows $p \to \infty$ provided that $p/n \to 0$, as in this case the condition of the theorem is satisfied.

OLS Robustness Under Model 1

In simple words: for "well-behaved" data, if $k/n \to 0$, OLS is robust to k sample removals, regardless of model mis-specification, and regardless of data dimension.

In fact, theorem allows $p \to \infty$ provided that $p/n \to 0$, as in this case the condition of the theorem is satisfied.

Robustness measure $\Delta_k(\boldsymbol{v})$ converges to zero, uniformly in $\boldsymbol{v}$:

$$\Delta_k(\boldsymbol{v}) = \max_{\mathcal{S}\subseteq[n],|\mathcal{S}|=n-k} \langle \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}, \boldsymbol{v}\rangle \leq \max_{\mathcal{S}\subseteq[n],|\mathcal{S}|\geq n-k} \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}\|.$$

Hence, if $k/n \to 0$,

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \Delta_k(\boldsymbol{v}) \xrightarrow{p} 0.$$

Consistency under sample removals / Model 1

Theorem

Under same conditions above, and additional assumption that $\kappa(\Sigma)$ *and* $\|\beta\|$ *remain bounded, as* $n \to \infty$ *and* $(p+k)/n \to 0$,

$$\max_{\mathcal{S} \subseteq [n], |\mathcal{S}| \geq n-k} \|\widehat{\beta}_{\mathcal{S}} - \beta\| \xrightarrow{p} 0.$$

Consistency under sample removals / Model 1

Theorem

Under same conditions above, and additional assumption that $\kappa(\Sigma)$ and $\|\beta\|$ remain bounded, as $n \to \infty$ and $(p+k)/n \to 0$,

$$\max_{\mathcal{S}\subseteq[n],|\mathcal{S}|\geq n-k} \|\widehat{\beta}_{\mathcal{S}} - \beta\| \xrightarrow{p} 0.$$

Maximum number of samples that can be removed while keeping OLS error rate:

$$\boxed{k \log\left(\frac{n}{k}\right) \ll \sqrt{p}}$$

Robustness / Model 2

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ i.i.d. from Gaussian Linear Model 2.

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ i.i.d. from Gaussian Linear Model 2.

Theorem

There exist constants $C, c > 0$ such that if

$$Ck \leq n, \quad t \geq 0, \quad \text{and} \quad \sqrt{\frac{p}{n-k}} + t < c,$$

then with probability at least $1 - 4e^{-c(n-k)t^2}$,

$$\max_{\substack{\mathcal{S} \subseteq [n], \\ |\mathcal{S}| \geq n-k}} \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}\| \leq C \|\Sigma^{-1/2}\| \|\varepsilon\|_{\psi_2} \left(\frac{k \log n}{n-k} + \sqrt{\frac{p}{n-k}} + t \right)$$

We conjecture theorem is optimal up to logarithmic factors.

Robustness under Linear Model

We conjecture theorem is optimal up to logarithmic factors.

The error rate of $\max_{\mathcal{S}\subseteq[n],|\mathcal{S}|\geq n-k}\|\widehat{\beta}_{\mathcal{S}}-\beta\|$ matches that of OLS on the full dataset if

$$k \ll \frac{\sqrt{np}}{\log n}$$

Robustness under Linear Model

We conjecture theorem is optimal up to logarithmic factors.

The error rate of $\max_{\mathcal{S}\subseteq[n],|\mathcal{S}|\geq n-k} \|\widehat{\beta}_{\mathcal{S}} - \beta\|$ matches that of OLS on the full dataset if

$$k \ll \frac{\sqrt{np}}{\log n}$$

Under linear model, OLS can tolerate the removal of significantly more samples than under the general model

On Theoretical Guarantees for ACRE

[Rubinstein and Hopkins, ICLR, 25']

ACRE= Algorithm for Certifying Robustness Efficiently

On Theoretical Guarantees for ACRE

[Rubinstein and Hopkins, ICLR, 25']

ACRE= Algorithm for Certifying Robustness Efficiently

For a fixed $\boldsymbol{v}$, ACRE computes upper and lower bounds $U_k(\boldsymbol{v})$ and $L_k(\boldsymbol{v})$ such that without any modeling assumptions,

$$L_k(\boldsymbol{v}) \leq \Delta_k(\boldsymbol{v}) \leq U_k(\boldsymbol{v}).$$

On Theoretical Guarantees for ACRE

Rubinstein and Hopkins derived following theoretical guarantee:

On Theoretical Guarantees for ACRE

Rubinstein and Hopkins derived following theoretical guarantee:

Under a variant of Model 2, with less restrictive conditions, there exists a threshold

$$K = \widetilde{\Theta}\left(\min \left(\frac{n}{\sqrt{p}}, \frac{n^2}{p^2} \right) \right)$$

such that for all $k \leq K$, with high probability,

$$\frac{U_k(\boldsymbol{v})}{L_k(\boldsymbol{v})} = 1 + \widetilde{O}\left(\frac{p + k\sqrt{p}}{n} \right).$$

On Theoretical Guarantees for ACRE

Rubinstein and Hopkins derived following theoretical guarantee:

Under a variant of Model 2, with less restrictive conditions, there exists a threshold

$$K = \widetilde{\Theta}\left(\min\left(\frac{n}{\sqrt{p}}, \frac{n^2}{p^2} \right) \right)$$

such that for all $k \leq K$, with high probability,

$$\frac{U_k(\boldsymbol{v})}{L_k(\boldsymbol{v})} = 1 + \widetilde{O}\left(\frac{p + k\sqrt{p}}{n} \right).$$

When $p + k\sqrt{p} \ll n$, the upper and lower bounds $U_k(\boldsymbol{v})$ and $L_k(\boldsymbol{v})$ are tight, so ACRE accurately measures robustness to removals.

On Theoretical Guarantees for ACRE

Comparison to our theoretical results:

On Theoretical Guarantees for ACRE

Comparison to our theoretical results:

Under model 2, OLS is *provably* robust to removals in a *broader parameter regime* $p + k \ll n$.

On Theoretical Guarantees for ACRE

Comparison to our theoretical results:

Under model 2, OLS is *provably* robust to removals in a *broader parameter regime* $p + k \ll n$.

Open Question: whether the upper and lower bounds of ACRE remain tight in more general regimes, in particular where OLS is non-robust.

Proof Techniques

For the general model 1:

Proof Techniques

For the general model 1:

proof is “standard”: based on concentration inequalities and union bounds.

Proof Techniques

For the general model 1:

proof is "standard": based on concentration inequalities and union bounds.

For Gaussian linear model 2:

Proof Techniques

For the general model 1:

proof is “standard”: based on concentration inequalities and union bounds.

For Gaussian linear model 2:

sharper results require more involved proof.

Proof Techniques

For the general model 1:

proof is "standard": based on concentration inequalities and union bounds.

For Gaussian linear model 2:

sharper results require more involved proof.

Careful use of Gaussian comparison inequalities

Proof Techniques

For the general model 1:

proof is "standard": based on concentration inequalities and union bounds.

For Gaussian linear model 2:

sharper results require more involved proof.

Careful use of Gaussian comparison inequalities

Open Question:

Proof Techniques

For the general model 1:

proof is "standard": based on concentration inequalities and union bounds.

For Gaussian linear model 2:

sharper results require more involved proof.

Careful use of Gaussian comparison inequalities

Open Question:

Derive robustness guarantees for other models

Non-robustness for $k \propto n$

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ i.i.d. from Model 2. Set $\alpha = k/n$.

Non-robustness for $k \propto n$

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ i.i.d. from Model 2. Set $\alpha = k/n$.

Theorem

Fix $\boldsymbol{v} \in \mathbb{S}^{p-1}$. Assume that $p < k$ and $\gamma = p/(n-k) < 1/4$. There exist absolute constants $C, c > 0$ such that if $\alpha = k/n \leq c$, then with probability at least $1 - 17e^{-c(n-k)t^2}$

$$\Delta_k(\boldsymbol{v}) \geq \left\|\Sigma^{-1/2}\boldsymbol{v}\right\| \left(\mathbb{E}[\varepsilon z \, \mathbb{1}(\varepsilon z > q_{1-\alpha+t})] - \frac{C(t+\sqrt{\gamma})}{(C - \sqrt{\gamma} - t)^2} \right),$$

for any $t \in (0, \min\{\alpha, 1/2 - \alpha\})$. Here, $z \sim \mathcal{N}(0,1)$ is independent of ε, and $q_{1-\alpha+t}$ is the $(1-\alpha+t)-$quantile of εz.

Robustness and Consistency Regimes for OLS / Model 2

Region	$k, p \ll n$	$k \ll n,\ p \asymp n$	$k \asymp n,\ p \ll n$	$k \asymp p \asymp n$
Robust	✓	✓	×	×
Consistent	✓	×	✓	×

Cash Transfers Dataset Revisited

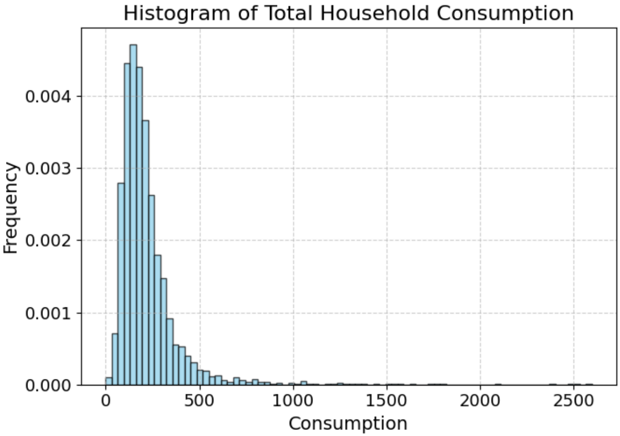
Response $Y =$ total household consumption in pesos

Cash Transfers Dataset Revisited

Response Y = total household consumption in pesos
In all 6 datasets, $\mathbb{E}[Y] \approx 200$, and std($Y$) comparable to $\mathbb{E}[Y]$

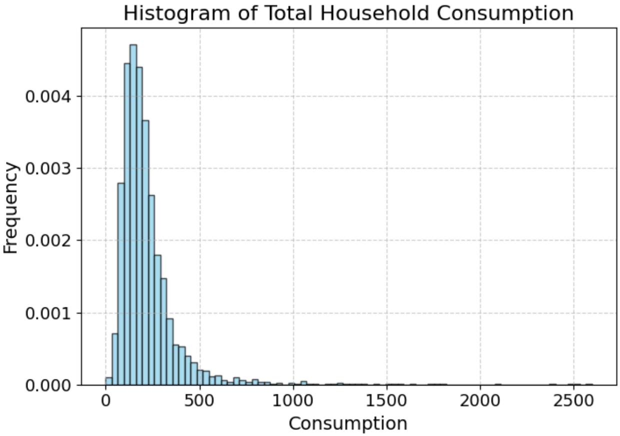
Cash Transfers Dataset Revisited

Response Y = total household consumption in pesos
In all 6 datasets, $\mathbb{E}[Y] \approx 200$, and std($Y$) comparable to $\mathbb{E}[Y]$



Cash Transfers Dataset Revisited

Response Y = total household consumption in pesos
In all 6 datasets, $\mathbb{E}[Y] \approx 200$, and std($Y$) comparable to $\mathbb{E}[Y]$



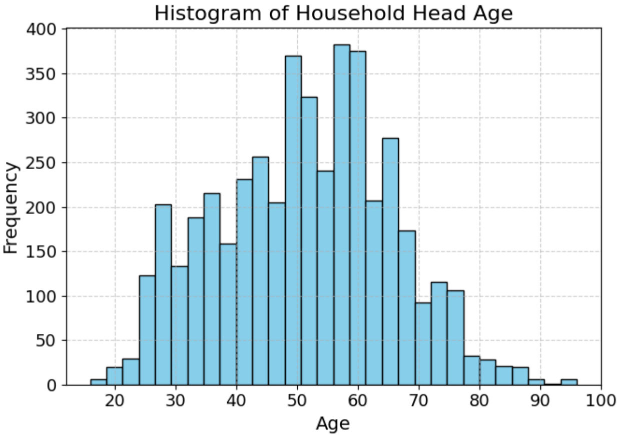
Y is heavy tailed

Cash Transfers Dataset

Explanatory variables $x_1, \ldots, x_{17}$ are all "well-behaved", some are categorical,

Cash Transfers Dataset

Explanatory variables $x_1, \ldots, x_{17}$ are all "well-behaved", some are categorical,



Cash Transfers Dataset

Period	Poor	n	$\widehat{\beta}_1$	AMIP	μ_y	σ_y	$>5\sigma_y$	$>10\sigma_y$
8	Y	10781	16.53	225	170	126	48	12
8	N	4543	-5.53	5	219	172	29	5
9	Y	9489	28.65	321	176	182	48	15
9	N	3769	23.19	21	226	273	20	9
10	Y	10368	32.52	570	172	156	56	13
10	N	4191	21.12	26	217	267	19	7

μ_y and σ_y: empirical mean and standard deviation of response y.
Last 2 columns: # samples larger than μ_y by $> 5\sigma_y$ or $> 10\sigma_y$

Cash Transfers Dataset

Period	Poor	n	$\widehat{\beta}_1$	AMIP	μ_y	σ_y	$>5\sigma_y$	$>10\sigma_y$
8	Y	10781	16.53	225	170	126	48	12
8	N	4543	-5.53	5	219	172	29	5
9	Y	9489	28.65	321	176	182	48	15
9	N	3769	23.19	21	226	273	20	9
10	Y	10368	32.52	570	172	156	56	13
10	N	4191	21.12	26	217	267	19	7

μ_y and σ_y: empirical mean and standard deviation of response y.
Last 2 columns: # samples larger than μ_y by $> 5\sigma_y$ or $> 10\sigma_y$

Y is *extremely heavy tailed*

Which samples are removed by AMIP ?

Which samples are removed by AMIP ?

Period	Poor	n	$\widehat{\beta}_1$	AMIP	μ_y	$\mu_{y^{\text{AMIP}}}$	$y_{\max}^{\text{AMIP}}$
8	Y	10781	16.53	225	170	572	4380
8	N	4543	-5.53	5	219	2018	2483
9	Y	9489	28.65	321	176	580	5117
9	N	3769	23.19	21	226	2670	5801
10	Y	10368	32.52	570	172	412	5080
10	N	4191	21.12	26	217	2154	7470

Which samples are removed by AMIP ?

Period	Poor	n	$\widehat{\beta}_1$	AMIP	μ_y	$\mu_{y^{\text{AMIP}}}$	$y_{\max}^{\text{AMIP}}$
8	Y	10781	16.53	225	170	572	4380
8	N	4543	-5.53	5	219	2018	2483
9	Y	9489	28.65	321	176	580	5117
9	N	3769	23.19	21	226	2670	5801
10	Y	10368	32.52	570	172	412	5080
10	N	4191	21.12	26	217	2154	7470

AMIP removes samples with extreme Y values

Which samples are removed by AMIP ?

Period	Poor	n	$\widehat{\beta}_1$	AMIP	μ_y	$\mu_{y^{\text{AMIP}}}$	$y_{\max}^{\text{AMIP}}$
8	Y	10781	16.53	225	170	572	4380
8	N	4543	-5.53	5	219	2018	2483
9	Y	9489	28.65	321	176	580	5117
9	N	3769	23.19	21	226	2670	5801
10	Y	10368	32.52	570	172	412	5080
10	N	4191	21.12	26	217	2154	7470

AMIP removes samples with extreme Y values

Perhaps not surprising few samples suffice to reverse sign $\hat{\beta}_1$

Robustness under Huber Loss

Suppose instead of OLS, we fit linear model under Huber loss

Robustness under Huber Loss

Suppose instead of OLS, we fit linear model under Huber loss

$$\widehat{\boldsymbol{\beta}}^{\text{Huber}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} h_\tau(\boldsymbol{\beta}^\top \mathbf{x}_i - y_i)$$

Robustness under Huber Loss

Suppose instead of OLS, we fit linear model under Huber loss

$$\widehat{\boldsymbol{\beta}}^{\text{Huber}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} h_\tau(\boldsymbol{\beta}^\top \mathbf{x}_i - y_i)$$

where h_τ is the Huber loss function, given by

$$h_\tau(z) = \begin{cases} \dfrac{z^2}{2} & |z| \leq \tau, \\ \tau\left(|z| - \dfrac{\tau}{2}\right) & |z| > \tau. \end{cases}$$

Robustness under Huber Loss

Suppose instead of OLS, we fit linear model under Huber loss

$$\widehat{\boldsymbol{\beta}}^{\text{Huber}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} h_\tau(\boldsymbol{\beta}^\top \mathbf{x}_i - y_i)$$

where h_τ is the Huber loss function, given by

$$h_\tau(z) = \begin{cases} \dfrac{z^2}{2} & |z| \leq \tau, \\ \tau\left(|z| - \dfrac{\tau}{2}\right) & |z| > \tau. \end{cases}$$

$\tau > 0$ controls the transition from squared loss to absolute loss.
In our experiments we took $\tau = 1$.

Robustness of Huber vs. OLS

Period	Poor	n	$\widehat{\beta}_1$	AMIP	$\widehat{\beta}_1^{\text{Huber}}$	AMIP Huber
8	Y	10781	16.53	225	16.55	725
8	N	4543	-5.53	5	-5.53	30
9	Y	9489	28.65	321	27.92	915
9	N	3769	23.19	21	22.15	228
10	Y	10368	32.52	570	31.31	1242
10	N	4191	21.12	26	19.06	217

Robustness of Huber vs. OLS

Period	Poor	n	$\widehat{\beta}_1$	AMIP	$\widehat{\beta}_1^{\text{Huber}}$	AMIP Huber
8	Y	10781	16.53	225	16.55	725
8	N	4543	-5.53	5	-5.53	30
9	Y	9489	28.65	321	27.92	915
9	N	3769	23.19	21	22.15	228
10	Y	10368	32.52	570	31.31	1242
10	N	4191	21.12	26	19.06	217

$\widehat{\beta}_1$ and $\widehat{\beta}_1^{\text{Huber}}$ -treatment effect under OLS and Huber regression

Robustness of Huber vs. OLS

Period	Poor	n	$\widehat{\beta}_1$	AMIP	$\widehat{\beta}_1^{\text{Huber}}$	AMIP Huber
8	Y	10781	16.53	225	16.55	725
8	N	4543	-5.53	5	-5.53	30
9	Y	9489	28.65	321	27.92	915
9	N	3769	23.19	21	22.15	228
10	Y	10368	32.52	570	31.31	1242
10	N	4191	21.12	26	19.06	217

$\widehat{\beta}_1$ and $\widehat{\beta}_1^{\text{Huber}}$ -treatment effect under OLS and Huber regression

Columns "AMIP" and "AMIP Huber": size of the smallest subset identified by AMIP whose removal reverses the sign of $\widehat{\beta}_1$ under each method.

Robustness of Huber vs. OLS

Period	Poor	n	$\widehat{\beta}_1$	AMIP	$\widehat{\beta}_1^{\text{Huber}}$	AMIP Huber
8	Y	10781	16.53	225	16.55	725
8	N	4543	-5.53	5	-5.53	30
9	Y	9489	28.65	321	27.92	915
9	N	3769	23.19	21	22.15	228
10	Y	10368	32.52	570	31.31	1242
10	N	4191	21.12	26	19.06	217

$\widehat{\beta}_1$ and $\widehat{\beta}_1^{\text{Huber}}$ -treatment effect under OLS and Huber regression

Columns “AMIP” and “AMIP Huber”: size of the smallest subset identified by AMIP whose removal reverses the sign of $\widehat{\beta}_1$ under each method.

Assuming AMIP approximation is accurate
Huber regression substantially more robust

Summary

Summary

- Robustness Auditing: important to enhance trust in a learned model. Framework goes beyond influence function of individual samples.

Summary

- Robustness Auditing: important to enhance trust in a learned model. Framework goes beyond influence function of individual samples.

- Presented theoretical analysis of robustness auditing for OLS.

Summary

- Robustness Auditing: important to enhance trust in a learned model. Framework goes beyond influence function of individual samples.

- Presented theoretical analysis of robustness auditing for OLS.

- Well behaved data and $k \ll n$, OLS is provably robust to sample removals.

Summary

- Robustness Auditing: important to enhance trust in a learned model. Framework goes beyond influence function of individual samples.

- Presented theoretical analysis of robustness auditing for OLS.

- Well behaved data and $k \ll n$, OLS is provably robust to sample removals.

- Implications: If removal of $k \ll n$ samples significantly changes linear model: need to carefully inspect potential reasons: heavy tails, outliers, non-i.i.d. data, etc.

Summary

- Robustness Auditing: important to enhance trust in a learned model. Framework goes beyond influence function of individual samples.

- Presented theoretical analysis of robustness auditing for OLS.

- Well behaved data and $k \ll n$, OLS is provably robust to sample removals.

- Implications: If removal of $k \ll n$ samples significantly changes linear model: need to carefully inspect potential reasons: heavy tails, outliers, non-i.i.d. data, etc.

- Multiple future directions: other models, less restrictive assumptions, heavy tailed distributions, outliers, etc.

Summary

- Robustness Auditing: important to enhance trust in a learned model. Framework goes beyond influence function of individual samples.

- Presented theoretical analysis of robustness auditing for OLS.

- Well behaved data and $k \ll n$, OLS is provably robust to sample removals.

- Implications: If removal of $k \ll n$ samples significantly changes linear model: need to carefully inspect potential reasons: heavy tails, outliers, non-i.i.d. data, etc.

- Multiple future directions: other models, less restrictive assumptions, heavy tailed distributions, outliers, etc.

Thank You !