Online Supplement to "*Integrating Data Cleaning and Robustification: A Study on Linear Regression*"

Utku Tarık Bilgiç, Xiaoning Qian, Bo Zeng

Appendix A: Proofs of Propositions and Theorems

The following proofs are derived based on the definitions and assumptions presented in Eq. (2) and Eq. (3).

A.1. Proof of Proposition 1

Let $\boldsymbol{\beta}_a^*$ and $\boldsymbol{\beta}_b^*$ be the optimal solutions to the minimization problems on training sets D_a and D_b, respectively. Let $q = \frac{|N_a^o|}{n} \leq \frac{1}{2}$,

$$\begin{aligned}
\mathbb{E}\big[\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}; D_a)\big] &= \mathbb{E}[\mathcal{L}(\boldsymbol{\beta}_a^*; D_a)] \\
&= \mathbb{E}[\sum_{i \in N_a} \eta(r_i(\boldsymbol{\beta}_a^*))] \\
&= \mathbb{E}[\sum_{i \in N_a^t} \eta(r_i(\boldsymbol{\beta}_a^*))] + \mathbb{E}[\sum_{i \in N_a^o} \eta(r_i(\boldsymbol{\beta}_a^*))] \\
&= \sum_{i \in N_a^t} \mathbb{E}[\eta(r_i(\boldsymbol{\beta}_a^*))] + \sum_{i \in N_a^o} \mathbb{E}[\eta(r_i(\boldsymbol{\beta}_a^*))] \\
&\geq (1-q)n\mathbb{E}[\eta(|\varepsilon|)] + qn\mathbb{E}[\eta(|\varepsilon^o|)].
\end{aligned}$$

While,

$$\mathbb{E}\big[\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}; D_b)\big] = \mathbb{E}[\mathcal{L}(\boldsymbol{\beta}_b^*; D_b)] = \mathbb{E}[\sum_{i \in N_b} \eta(r_i(\boldsymbol{\beta}_b^*))] = \sum_{i \in N_b} \mathbb{E}[\eta(r_i(\boldsymbol{\beta}_b^*))] = n\mathbb{E}[\eta(|\varepsilon|)].$$

As defined in Eq. (3), $\mathbb{E}[\eta(|\varepsilon^o|)] \geq \mathbb{E}[\eta(|\varepsilon|)]$ and,

$$\mathbb{E}\big[\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}; D_a)\big] \geq (1-q)n\mathbb{E}[\eta(|\varepsilon|)] + qn\mathbb{E}[\eta(|\varepsilon^o|)] \geq n\mathbb{E}[\eta(|\varepsilon|)] = \mathbb{E}\big[\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}; D_b)\big]. \quad \square$$

A.2. Proof of Theorem 1

Let $D^{\boldsymbol{w}} = \{(\boldsymbol{x}_i, y_i) | w_i = 1, i \in N\}$ and $N^{\boldsymbol{w}}$ the associated index set. Let $\boldsymbol{w}^*$ be an optimal solution to the left hand side (LHS) of Eq. (7).

$\Leftarrow$

Considering $\boldsymbol{w} \in \mathcal{Z}_{l^*}^D$, we have

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^{\boldsymbol{w}^*}}} \boldsymbol{z}^T |\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}| \leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^{\boldsymbol{w}}}} \boldsymbol{z}^T |\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}| \\
&\mathbb{E}\Big[\min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^{\boldsymbol{w}^*}}} \boldsymbol{z}^T |\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}|\Big] \leq \mathbb{E}\Big[\min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^t}} \boldsymbol{z}^T |\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}|\Big].
\end{aligned}$$

$\Rightarrow$

$D^{\boldsymbol{w}} = D_1^{\boldsymbol{w}} \cup D_0^{\boldsymbol{w}}$, where $D_1^{\boldsymbol{w}} = \{(\boldsymbol{x}_i, y_i) | w_i = 1, z_i = 1,\ i \in N\}$ and $D_0^{\boldsymbol{w}} = \{(\boldsymbol{x}_i, y_i) | w_i = 1, z_i = 0,\ i \in N\}$, with index sets $N_1^{\boldsymbol{w}}$ and $N_0^{\boldsymbol{w}}$, respectively. As a higher level decision, all decision variables are dependent on the set of points chosen by $\boldsymbol{w}$. Let $(\boldsymbol{w}^*, \boldsymbol{\beta}^*, \boldsymbol{z}^*)$ and $(\boldsymbol{\beta}^\dagger, \boldsymbol{z}^\dagger)$ be optimal solutions for LHS and RHS, respectively. $|D_1^{w^*}| = k$

and $|D_0^{w^*}| = l^* - k$.

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^{\boldsymbol{w}^*}}} \sum_{i=1}^{n} z_i |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}| = \sum_{i \in N_1^{\boldsymbol{w}^*}} |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^*|.$$

Similarly for RHS:

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^t}} \sum_{i=1}^{n} z_i |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}| = \sum_{i \in N_{t,1}} |y_i - x_i^T \boldsymbol{\beta}^\dagger|.$$

Let $r_i^* = |y_i - \boldsymbol{x}_i \boldsymbol{\beta}^*|$,

(i) If $D_1^{\boldsymbol{w}^*} \neq D_1^t$, then $D_1^{\boldsymbol{w}^*} \subset D^t \cup D^o$.

Since $D_1^t \subset D^t$, $\mathbb{E}[\boldsymbol{\beta}^\dagger] = \mathbb{E}[\boldsymbol{\beta}^t]$, and $\mathbb{E}[r_i^\dagger] = \mathbb{E}[|\varepsilon_i|]$, $\forall i \in D_1^t$:

$$\mathbb{E}\left[\sum_{i \in \{N_1^{\boldsymbol{w}^*} \cap N^t\}} r_i^* + \sum_{i \in \{N_1^{\boldsymbol{w}^*} \cap N^o\}} r_i^*\right] \geq \mathbb{E}\left[\sum_{i \in \{N_1^{\boldsymbol{w}^*} \cap N^t\}} |\varepsilon_i| + \sum_{i \in \{N_1^{\boldsymbol{w}^*} \cap N^o\}} |\varepsilon_i^o|\right] \geq \mathbb{E}\left[\sum_{i \in N_1^t} |\varepsilon_i|\right] = \mathbb{E}\left[\sum_{i \in N_1^t} r_i^\dagger\right].$$

(ii) If $D^{\boldsymbol{w}^*} = D^t$, then $\mathbb{E}[\boldsymbol{\beta}^*] = \mathbb{E}[\boldsymbol{\beta}^\dagger] = \mathbb{E}[\boldsymbol{\beta}^t]$:

$$\sum_{i \in N_1^{\boldsymbol{w}^*}} \mathbb{E}[r_i^*] = \sum_{i \in N_1^t} \mathbb{E}[r_i^\dagger],$$

therefore,

$$\mathbb{E}\left[\min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^{\boldsymbol{w}^*}}} \boldsymbol{z}^T |\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}|\right] \geq \mathbb{E}\left[\min_{\boldsymbol{\beta}} \max_{\boldsymbol{z} \in \mathcal{Z}_k^{D^{\boldsymbol{w}^*}}} \boldsymbol{z}^T |\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}|\right]. \quad \square$$

A.3. Proof of Theorem 2

Wasserstein distance between the outlier distribution and sample distributions, $W_1(P_o, \mathcal{P}_D)$, can be defined as (Chen and Paschalidis (2020), *Theorem 2.6.1*)

$$W_1(P_o, \mathcal{P}_D) = W_1(P_o, P_t) - W_1(P_t, \mathcal{P}_D) = (1-q) W_1(P_o, P_t).$$

$$W_1(P_o, \mathcal{P}_{D'}) = W_1(P_o, P_t) - W_1(P_t, \mathcal{P}_{D'}) = (1-q') W_1(P_o, P_t).$$

For $q' < q$, then

$$W_1(P_o, \mathcal{P}_{D'}) > W_1(P_o, \mathcal{P}_D) \text{ and } W_1(P_t, \mathcal{P}_{D'}) < W_1(P_t, \mathcal{P}_D). \quad \square$$

A.4. Proof of Theorem 3

Let $\mathcal{M}(\Xi)$ be the space of all probability distributions Q supported on Ξ. Q_i is the marginal conditional distribution given sample $(\boldsymbol{x}_i, y_i)$. The inner problem in the first row of Eq. (11), which denotes the worst-case expectation for the chosen subset (S_l), can be expressed as

$$\begin{aligned}
\sup_{Q \in \mathbb{B}(\mathcal{P}_{S_l})} \mathbb{E}^Q\left[|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}|\right] = & \sup_{Q_i \in \mathcal{M}(\mathbb{R}^{p+1})} \frac{1}{l} \sum_{i \in N^{S_l}} \int_{\mathbb{R}^{p+1}} |y - \boldsymbol{x}^T \boldsymbol{\beta}| Q_i(d(\boldsymbol{x}, y)) \\
& s.t. \ \frac{1}{l} \sum_{i \in N^{S_l}} \int_{\mathbb{R}^{p+1}} |(\boldsymbol{x}, y) - (\boldsymbol{x}_i, y_i)| Q_i(d(\boldsymbol{x}, y)) \leq \epsilon \\
\leq & \inf_{\lambda \geq 0} \sup_{Q_i \in \mathcal{M}(\mathbb{R}^{p+1})} \frac{1}{l} \sum_{i \in N^{S_l}} \int_{\mathbb{R}^{p+1}} |y - \boldsymbol{x}^T \boldsymbol{\beta}| Q_i(d(\boldsymbol{x}, y)) \qquad \text{(A1)}
\end{aligned}$$

$$+\lambda(\epsilon-\frac{1}{l}\sum_{i\in N^{S_l}}\int_{\mathbb{R}^{p+1}}|(\boldsymbol{x},y)-(\boldsymbol{x}_i,y_i)|Q_i(d(\boldsymbol{x},y)))$$
$$=\inf_{\lambda\geq 0}\lambda\epsilon+\frac{1}{l}\sum_{i\in N^{S_l}}\sup_{(\boldsymbol{x},y)\in\mathbb{R}^{p+1}}(|y-\boldsymbol{x}^T\boldsymbol{\beta}|-\lambda|(\boldsymbol{x},y)-(\boldsymbol{x}_i,y_i)|).$$

For $\epsilon\geq 0$, the inequality Eq. (A1) is exact due to strong duality of the generalized moment problem (Shapiro 2001, Esfahani and Kuhn 2018). Since LAD is proper, convex and lower semicontinuous,

$$\sup_{(\boldsymbol{x},y)\in\mathbb{R}^{p+1}}(|y-\boldsymbol{x}^T\boldsymbol{\beta}|-\lambda|(\boldsymbol{x},y)-(\boldsymbol{x}_i,y_i)|)\leq\begin{cases}|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}|, & \textit{if } \sup\{|\boldsymbol{\theta}|_\infty:\boldsymbol{\theta}\in\Theta\}\leq\lambda\\ \infty, & \textit{otherwise}\end{cases},$$

where $\Theta:=\{\boldsymbol{\theta}\in\mathbb{R}^{p+1}:\mathcal{L}^*(\boldsymbol{\theta})<\infty\}$ and the conjugate of the absolute value loss function $\mathcal{L}^*(\boldsymbol{\theta})=\sup_{(\boldsymbol{x},y)\in\mathbb{R}^{p+1}}(\boldsymbol{x},y)^T\boldsymbol{\theta}-|y-\boldsymbol{x}^T\boldsymbol{\beta}|$.

$$\mathcal{L}^*(\boldsymbol{\theta})=\begin{cases}0, & \textit{if } (\boldsymbol{x},y)^T\boldsymbol{\theta}=|y-\boldsymbol{x}^T\boldsymbol{\beta}|\\ \infty, & \textit{otherwise}\end{cases}$$

Since $\boldsymbol{\theta}\in\Theta$, $\sup\{|\boldsymbol{\theta}|_\infty:\boldsymbol{\theta}\in\Theta\}=|(-\boldsymbol{\beta},1)|_\infty$ (Chen and Paschalidis (2018), *Appendix A.1*). This upper-bound becomes exact since $\Xi=\mathbb{R}^{p+1}$ (Esfahani and Kuhn (2018), *Theorem 6.3*). Then, we can write the worst-case expectation as

$$\inf_{|(\boldsymbol{\beta},1)|_\infty\leq\lambda}\lambda\epsilon+\frac{1}{l}\sum_{i\in N^{S_l}}|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}|=\epsilon|(\boldsymbol{\beta},1)|_\infty+\frac{1}{l}\sum_{i\in N^{S_l}}|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}|.$$

The subpopulation S_l is defined by the decision of either including or eliminating the data point in the model training, via $\boldsymbol{w}\in\{0,1\}$. Constraint $|\boldsymbol{w}|_0=l$ is added to assure the size of the S_l. Then the single-level formulation is

$$\min_{\substack{\boldsymbol{w}\in\{0,1\}^n\\|\boldsymbol{w}|_0=l}}\min_{\boldsymbol{\beta}}\epsilon|(\boldsymbol{\beta},1)|_\infty+\sum_{i=1}^n w_i|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}|.\quad\square$$

A.5. Proof of Theorem 4

For any given ϵ, $J^*_{S_l}\leq J^*_D$ since the subpopulation is achieved by eliminating outliers which are data points with high residuals. From Theorem 2, we know $W_1(P_t,\mathcal{P}_{S_l})\leq\epsilon'<W_1(P_t,\mathcal{P}_D)\leq\epsilon$,

$$P(W_1(P_t,\mathcal{P}_{S_l})\geq\epsilon)\leq\alpha'\leq P(W_1(P_t,\mathcal{P}_D)\geq\epsilon)\leq\alpha,$$

and then,

$$P(\mathbb{E}[\mathcal{L}(\boldsymbol{\beta}^*_{S_l};T)]\leq J^*_{S_l})\geq P(P_t\in\mathbb{B}(\mathcal{P}_{S_l}))$$
$$\geq 1-\alpha'.\quad\square$$

A.6. Proof of Theorem 5

Let $r_{[i]}(\boldsymbol{\beta})$ be the i-th largest residual for $i=1,\ldots,n$, where $0\leq r_{[1]}(\boldsymbol{\beta})\leq r_{[2]}(\boldsymbol{\beta})\leq\cdots\leq r_{[n]}(\boldsymbol{\beta})$. Then, $\mathcal{L}(\boldsymbol{w}^*_l,\boldsymbol{\beta}^*_l)=\sum_{i=1}^l\eta(r_{[i]}(\boldsymbol{\beta}^*_l))$.

CLAIM 1. *For any $l\geq\frac{n}{2}$ and $\boldsymbol{\beta}\neq\boldsymbol{\beta}^*_l$,*

$$\sum_{i=1}^l\eta(r_{[i]}(\boldsymbol{\beta}^*_l))\leq\sum_{j=1}^l\eta(r_{[j]}(\boldsymbol{\beta})).$$

Proof of Claim

$$\mathcal{L}(\boldsymbol{w}_l^*, \boldsymbol{\beta}_l^*) = \sum_{i=1}^{l} \eta(r_{[i]}(\boldsymbol{\beta}_l^*)) = \min_{\boldsymbol{w},\boldsymbol{\beta}} \sum_{i=1}^{n} w_i \eta(r_i(\boldsymbol{\beta}))$$
$$s.t.\ |\boldsymbol{w}|_0 = l$$
$$\boldsymbol{w} \in \{0,1\}^n$$

As any feasible $(\boldsymbol{w}, \boldsymbol{\beta}) \neq (\boldsymbol{w}_l^*, \boldsymbol{\beta}_l^*)$ *will lead to feasible solution for the minimization problem, the objective function value will be more than or equal to the objective function value of the optimal solution.* □

For $l^* > l''$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l''+1}) - \mathcal{L}(\boldsymbol{w}^*_{l''})\right] &= \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l''+1})\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l''})\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{l''+1} \eta(r_{[i]}(\boldsymbol{\beta}^*_{l''+1}))\right] - \mathbb{E}\left[\sum_{j=1}^{l''} \eta(r_{[j]}(\boldsymbol{\beta}^*_{l''}))\right] \\
&= \mathbb{E}[\eta(r_{[l''+1]}(\boldsymbol{\beta}^*_{l''+1}))] + \mathbb{E}\left[\sum_{i=1}^{l''} \eta(r_{[i]}(\boldsymbol{\beta}^*_{l''+1}))\right] - \mathbb{E}\left[\sum_{j=1}^{l''} \eta(r_{[j]}(\boldsymbol{\beta}^*_{l''}))\right] \\
&= \mathbb{E}[\eta(|\varepsilon|)] + l''\big(\mathbb{E}[\eta(|\varepsilon|)] - \mathbb{E}[\eta(|\varepsilon|)]\big) \\
&= \mathbb{E}[\eta(|\varepsilon|)]
\end{aligned}$$

Since both model parameters $\boldsymbol{\beta}^*_{l''}$ and $\boldsymbol{\beta}^*_{l''+1}$ are estimated by training the model on samples only from the true distribution, $\mathbb{E}\left[\boldsymbol{\beta}^*_{l''}\right] = \mathbb{E}\left[\boldsymbol{\beta}^*_{l''+1}\right]$.

For $l' \geq l^*$, if $l' = l^*$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l^*+1}) - \mathcal{L}(\boldsymbol{w}^*_{l^*})\right] &= \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l^*+1})\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l^*})\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{l^*+1} \eta(r_{[i]}(\boldsymbol{\beta}^*_{l^*+1}))\right] - \mathbb{E}\left[\sum_{j=1}^{l^*} \eta(r_{[j]}(\boldsymbol{\beta}^*_{l^*}))\right] \\
&= \mathbb{E}\big[\eta(r_{[l'+1]}(\boldsymbol{\beta}^*_{l^*+1}))\big] + \sum_{i=1}^{l^*} \mathbb{E}\big[\eta(r_{[i]}(\boldsymbol{\beta}^*_{l^*+1}))\big] - \sum_{j=1}^{l^*} \mathbb{E}\big[\eta(r_{[j]}(\boldsymbol{\beta}^*_{l^*}))\big] \\
&\geq \mathbb{E}\big[\eta(|\varepsilon^o|)\big] + l^*\mathbb{E}\big[\eta(|\varepsilon|)\big] - l^*\mathbb{E}\big[\eta(|\varepsilon|)\big] \\
&\geq \mathbb{E}[\eta(|\varepsilon^o|)].
\end{aligned}$$

Otherwise if $l' > l^*$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l'+1}) - \mathcal{L}(\boldsymbol{w}^*_{l'})\right] &= \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l'+1})\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l'})\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{l'+1} \eta(r_{[i]}(\boldsymbol{\beta}^*_{l'+1}))\right] - \mathbb{E}\left[\sum_{j=1}^{l'} \eta(r_{[j]}(\boldsymbol{\beta}^*_{l'}))\right] \\
&= \mathbb{E}\Big[\eta(r_{[l'+1]}(\boldsymbol{\beta}^*_{l'+1}))\Big] + \mathbb{E}\Big[\sum_{i=1}^{l'} \eta(r_{[i]}(\boldsymbol{\beta}^*_{l'+1})) - \sum_{j=1}^{l'} \eta(r_{[j]}(\boldsymbol{\beta}^*_{l'}))\Big] \\
&\geq \mathbb{E}\Big[\eta(r_{[l'+1]}(\boldsymbol{\beta}^*_{l'+1}))\Big] \\
&\geq \mathbb{E}[\eta(|\varepsilon^o|)].
\end{aligned}$$

The first inequality comes from the difference between a feasible solution and the optimal solution, explained in Claim 1.

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l'+1}, \boldsymbol{\beta}^*_{l'+1}) - \mathcal{L}(\boldsymbol{w}^*_{l'}, \boldsymbol{\beta}^*_{l'})\right] \geq \mathbb{E}[\eta(|\varepsilon^o|)] \geq \mathbb{E}[\eta(|\varepsilon|)] = \mathbb{E}\left[\mathcal{L}(\boldsymbol{w}^*_{l''+1}, \boldsymbol{\beta}^*_{l''+1}) - \mathcal{L}(\boldsymbol{w}^*_{l''}, \boldsymbol{\beta}^*_{l''})\right] \quad \square$$

Appendix B: Algorithm Details and Extended Results

B.1. Estimating the Size of the Outlier

Algorithm A1 Estimating the Size of the Outlier Set

Initialize $L_{\frac{n}{2}} = \text{CALCLOSS}(\frac{n}{2})$

for $l = \frac{n}{2}+1, \ldots, n$ **do**

$L_l = \text{CALCLOSS}(l)$

$\Delta L_l = L_l - L_{l-1}$

end for

$\boldsymbol{l} = (\frac{n}{2}, \ldots, n)$

return $l^* = \arg\min_{l'} \left(\min_{\boldsymbol{\beta}} \sum_{l=\frac{n}{2}}^{l'} |\Delta L_l - \boldsymbol{l}^T \boldsymbol{\beta}_1| + \sum_{l=l'}^{n} |\Delta L_l - \boldsymbol{l}^T \boldsymbol{\beta}_2| \right)$

For small and mid-size data sets, it is relatively easy to solve models to optimally. For calculating the training loss (CalcLoss(l)), LTA (with optimality) or iterative LTA (with approximate results) can be used. Iterative LTA (itLTA) involves repeatedly selecting samples with the lowest current loss and re-fitting the linear model on those samples. However, in large data sets, solving the outlier elimination model many times is computationally heavy. Instead, one idea is to consider a small number of samples and solving for those samples optimally could be a quick and unbiased way of estimating the outlier proportion. Also, rather than the exact value of the training loss, the shape of the curve is critical for finding the size of outlier set. Hence, there is a flexibility in calculating the training loss either optimally or approximately. Depending on factors such as data set size, computational constraints, and user preferences, one could alternatively use exact or iterative LTA or sampling based algorithms. One more idea to reduce the computational burden is to first compute for a coarser grid and then a finer grid within the target interval.

B.2. ADMM Algorithm for C-DRO

The details of the algorithm are defined in Algorithm A2.

B.3. Hyperparameter Tuning

Section 5.1 For each model, we tuned the hyperparameters by training the models on separate validation sets with data and outliers generated similarly as described in Section 5.1.1. We generated 10 instances and performed grid search to identify the optimal model-specific parameters. For "Outliers in y";

- For k (for SR; $\%n$ and for C-RO; $\%l$): Equally spaced between $(0.5, 1.0)$ in increments of 0.05.
- For ϵ: Exponential values of the range equally spaced between $[-2.5, -0.5]$ in increments of 0.05. It is also scaled by the size of the training set.

For "Outliers in x and y", we only changed the range to $[-5, 0]$ for the ϵ

Algorithm A2 ADMM for C-DRO

Initialize $L_\rho^0 = 0$, $\boldsymbol{\theta}^0$, $\boldsymbol{w}$, $\boldsymbol{\lambda}^0 = \mathbf{0}$, $c = 0$, c', $\tau \in [1, 2]$ and $t = 1$

while $|L_\rho^t - L_\rho^{t-1}| \leq \delta_1$ & $|\lambda^t - \lambda^{t-1}| \leq \delta_2$ **do**

$\boldsymbol{\theta}^{t+1} := \arg\min_{\boldsymbol{\theta}} \ \boldsymbol{w}^t|\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}| + \boldsymbol{\lambda}^t(\boldsymbol{\beta}^t - \boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\beta}^t - \boldsymbol{\theta}\|_2^2$

$\boldsymbol{\beta}^{t+1} := \arg\min_{\boldsymbol{\beta}} \ \epsilon|(\boldsymbol{\beta}, 1)|_\infty + \boldsymbol{\lambda}^t(\boldsymbol{\beta} - \boldsymbol{\theta}^{t+1}) + \frac{\rho}{2}\|\boldsymbol{\beta} - \boldsymbol{\theta}^{t+1}\|_2^2$

$\boldsymbol{w}^{t+1} := \arg\min_{\boldsymbol{w} \in \mathcal{Z}_l} \ \boldsymbol{w}|Y - X\theta^{t+1}|$

$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \rho(\boldsymbol{\beta}^{t+1} - \boldsymbol{\theta}^{t+1})$

if $L_\rho^{t-1} < L_\rho^t$ **then** $c \leftarrow c + 1$

if $c \geq c'$ **then** $\rho \leftarrow \tau\rho$

end if

else $c \leftarrow 0$

end if

$t \leftarrow t + 1$

end while

Section 5.2 For each models and data sets, after training/test split, we sampled 40 points from the training set and generated the outliers similarly as described in Section 5.2. We used the remaining set as validation. We repeated this for 5 times and then performed grid search for the best hyperparameters.

- For k (for SR; $\%n$ and for C-RO; $\%l$): Equally spaced between $(0.5, 1.0)$ in increments of 0.05.
- For ϵ: Exponential values of the range equally spaced between $[-7, -1]$ in increments of 0.5. It is also scaled by the size of the training set.

Section 5.3

Black-box Attacks For each model, we sampled 40 points from the training set and replaced varying percentages with generated poisoning attacks based on the separate substitute data set. We repeated this for 5 times and then performed grid search for the best hyperparameters. The grid described above for hyperparameter tuning of Section 5.2 is used.

For C-RO and C-DRO, two different values of l ($l = 1 - \%$ poisoning, $l = 0.85 - \%$ poisoning) are used. In Table 2, results of best performing l value is reported.

White-box Attack For each model, we sampled 100 points from the training set and divided it into $80\%/20\%$ training/test split. We generated poisoning points based on varying percentages with Nopt, on top of the training set specific for hyperparameter tuning. We repeated this for 5 times and then performed grid search for the best hyperparameters.

- For k (for SR; $\%n$ and for C-RO; $\%l$): Equally spaced between $(0.5, 1.0)$ in increments of 0.05.
- For ϵ: Exponential values of the range equally spaced between $[-8, -0]$ in increments of 0.25. It is also scaled by the size of the training set.

B.4. Synthetic Data with Outliers

Table A1 Average MSE and MAE Results and Other Performance Metrics of Models Trained with Outliers

	Outliers in y						Outliers in X and y					
	MSE	**MAE**	$\|\beta-\beta^*\|$	**% of outliers eliminated**	**k**	ϵ	**MSE**	**MAE**	$\|\beta-\beta^*\|$	**% of outliers eliminated**	**k**	ϵ
OLS	2.47	1.25	0.81	-	-	-	31.73	4.87	2.98	-	-	-
LAD	1.21	0.88	0.34	-	-	-	28.91	4.52	3.16	-	-	-
LTA	1.14	0.84	0.29	81%	-	-	1.13	0.85	0.30	90%	-	-
SR	1.20	0.87	0.33	-	85%	-	28.70	4.51	3.11	-	90%	-
C-RO	1.12	0.85	0.27	81%	55%	-	1.09	0.83	0.26	90%	60%	-
DRR	1.05	0.82	0.10	-	-	0.58	29.02	4.54	3.13	-	-	0.02
C-DRO	1.10	0.84	0.15	82%	-	0.30	1.05	0.82	0.14	90%	-	0.29

Average of 50 experiments are reported in the table. For the models with outlier elimination (LTA, C-RO, C-DRO), hyperparameter $l = 0.8 \times n$ is fixed. Details of hyperparameter tuning are reported in Appendix B.3.

B.5. Performance of the ADMM algorithm

Table A2 Performance and Solution Quality of the ADMM Algorithm with and without Neighborhood Search Compared to the Gurobi Solver

	Outliers in y			Outliers in x and y		
	# of times opt. soln. found	**Training Loss**		**# of times opt. soln. found**	**Training Loss**	
		Mean	**St. Dev.**		**Mean**	**St. Dev.**
C-RO		22.48	2.78		24.64	3.17
C-RO (ADMM)	15/50	23.04	2.86	0/50	107.47	34.7
C-RO (ADMM + NS)	50/50	22.48	2.78	50/50	24.64	3.17
C-DRO		52.09	3.71		52.29	3.99
C-DRO (ADMM)	45/50	52.2	3.69	37/50	65.43	36.15
C-DRO (ADMM + NS)	50/50	52.09	3.71	50/50	52.29	3.99

Mean and standard deviation are calculated by results of 50 experiments. C-RO and C-DRO are solved to optimality by Gurobi Solver with default optimality gap 10^{-4}. For ADMM and ADMM + NS, if the algorithm solution is within 1% of the true optimal calculated by Gurobi Solver, we count them in the "# of times opt. soln. found" column.

B.6. Poisoning Attacks: Implementation Details and Extended Results

White-box Attack For our experiments, while creating poisoning points via Nopt, we adopt the LAD loss in line with the rest of the paper. To improve our solution capacity to the involved bilevel optimization model and the overall computational efficiency of Nopt, a few techniques have been employed. We first derive gradient approximation by finite difference. Also, all features are scaled to the $[0, 1]$ range to improve numerical stability and reduce the variance in gradient approximation. Additionally, rather than updating all poisoned points and features in each iteration, we sample a subset of points and features and apply updates only to the selected entries.

Table A3 Average MSE and MAE Results and Eliminated Poisoned Samples (EPS) by Data Cleaning Scheme in our Proposed Models under Black-box Attacks.

		StatP												Flip									
		0.05			0.1			0.15			0.2			0.05			0.1			0.15			0.2		
		MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS
abalone	**OLS (no attack)**	5.21	1.65		5.21	1.65		5.21	1.65		5.21	1.65		5.21	1.65		5.21	1.65		5.21	1.65		5.21	1.65	
	LAD (no attack)	5.31	1.60		5.31	1.60		5.31	1.60		5.31	1.60		5.31	1.60		5.31	1.60		5.31	1.60		5.31	1.60	
	OLS	7.92	2.03		9.97	2.30		11.60	2.50		12.90	2.66		8.34	2.31		13.74	3.04		19.85	3.77		27.15	4.55	
	LAD	6.39	1.72		8.91	2.09		12.84	2.59		13.56	2.67		5.35	1.61		5.38	1.63		5.47	1.68		5.58	1.73	
	SR	6.34	1.72		10.15	2.31		11.75	2.49		12.62	2.58		5.29	1.64		5.49	1.71		5.48	1.70		5.57	1.73	
	DRR	6.31	1.68		7.89	1.86		8.95	2.00		10.33	2.18		5.74	1.64		5.67	1.65		5.66	1.68		5.70	1.73	
	C-RO (1-% pois.)	5.85	1.68	80%	9.88	2.24	32%	12.73	2.58	20%	13.08	2.61	22%	5.29	1.61	99%	5.27	1.61	99%	5.28	1.61	99%	5.35	1.61	100%
	C-RO (0.85-% pois.)	6.27	1.66	97%	7.18	1.81	88%	12.78	2.56	28%	13.90	2.70	23%	5.48	1.60	100%	5.60	1.60	100%	5.61	1.61	100%	5.44	1.61	100%
	C-DRO (1-% pois.)	5.60	1.62	98%	6.30	1.68	94%	7.73	1.83	83%	13.50	2.59	49%	5.76	1.63	99%	5.72	1.63	99%	5.72	1.63	99%	5.63	1.62	99%
	C-DRO (0.85-% pois.)	6.54	1.67	100%	13.83	2.64	54%	14.10	2.67	56%	14.51	2.69	50%	6.56	1.67	100%	6.65	1.68	100%	6.65	1.68	100%	6.71	1.68	100%
autompg	**OLS (no attack)**	13.13	2.64		13.13	2.64		13.13	2.64		13.13	2.64		13.13	2.64		13.13	2.64		13.13	2.64		13.13	2.64	
	LAD (no attack)	15.97	2.75		15.97	2.75		15.97	2.75		15.97	2.75		15.97	2.75		15.97	2.75		15.97	2.75		15.97	2.75	
	OLS	24.67	3.70		33.99	4.46		38.74	4.82		43.92	5.18		20.55	3.60		31.21	4.63		40.29	5.43		48.91	6.09	
	LAD	18.42	2.88		25.59	3.42		34.55	4.25		48.41	5.32		16.18	2.75		15.61	2.72		14.22	2.69		14.16	2.74	
	SR	17.94	2.85		25.99	3.44		34.59	4.29		43.76	5.13		14.17	2.64		15.07	2.68		14.29	2.69		14.01	2.75	
	DRR	18.42	2.88		21.85	3.27		24.05	3.61		27.67	4.07		15.90	2.75		15.30	2.72		14.03	2.69		13.92	2.74	
	C-RO (1-% pois.)	16.15	2.79	95%	17.43	2.93	83%	18.15	2.97	88%	18.49	2.99	87%	14.45	2.68	100%	14.36	2.68	94%	14.31	2.68	99%	14.40	2.69	100%
	C-RO (0.85-% pois.)	17.49	2.88	94%	19.62	3.02	96%	19.51	3.03	96%	19.29	3.01	96%	19.73	2.98	100%	18.71	2.95	100%	19.04	3.00	100%	18.87	2.99	100%
	C-DRO (1-% pois.)	16.03	2.76	100%	17.68	2.91	91%	29.93	3.92	43%	32.53	4.08	48%	16.03	2.76	100%	15.43	2.74	95%	15.57	2.74	100%	15.60	2.76	100%
	C-DRO (0.85-% pois.)	17.64	2.88	100%	19.87	3.07	92%	34.83	4.11	51%	54.65	5.39	35%	19.85	3.00	100%	17.43	2.87	100%	17.80	2.93	100%	18.50	2.96	100%
wine	**OLS (no attack)**	0.41	0.50		0.41	0.50		0.41	0.50		0.41	0.50		0.41	0.50		0.41	0.50		0.41	0.50		0.41	0.50	
	LAD (no attack)	0.42	0.50		0.42	0.50		0.42	0.50		0.42	0.50		0.42	0.50		0.42	0.50		0.42	0.50		0.42	0.50	
	OLS	0.47	0.55		0.49	0.56		0.51	0.57		0.53	0.58		0.45	0.53		0.54	0.59		0.59	0.63		0.66	0.68	
	LAD	0.46	0.54		0.51	0.57		0.57	0.60		0.58	0.60		0.42	0.50		0.43	0.51		0.43	0.51		0.44	0.52	
	SR	0.47	0.55		0.49	0.56		0.52	0.57		0.53	0.58		0.41	0.50		0.43	0.51		0.43	0.51		0.43	0.52	
	DRR	0.48	0.53		0.50	0.56		0.56	0.59		0.58	0.60		0.43	0.51		0.44	0.52		0.44	0.52		0.45	0.52	
	C-RO (1-% pois.)	0.48	0.54	16%	0.51	0.57	8%	0.55	0.59	4%	0.58	0.60	6%	0.42	0.50	92%	0.42	0.50	95%	0.42	0.51	96%	0.42	0.51	96%
	C-RO (0.85-% pois.)	0.53	0.56	19%	0.56	0.59	13%	0.66	0.64	6%	0.67	0.65	6%	0.43	0.51	99%	0.42	0.50	100%	0.42	0.51	100%	0.42	0.51	100%
	C-DRO (1-% pois.)	0.43	0.50	100%	0.43	0.50	100%	0.60	0.61	4%	0.62	0.62	5%	0.44	0.51	88%	0.44	0.51	94%	0.44	0.51	95%	0.44	0.51	96%
	C-DRO (0.85-% pois.)	0.44	0.51	100%	0.44	0.51	100%	1.84	0.94	6%	1.98	0.98	7%	0.44	0.51	100%	0.45	0.51	100%	0.45	0.51	100%	0.45	0.51	100%

Detailed results of average of 8 experiments are reported. The best regression model for each case is underlined and marked in boldface, and the second-best model is just marked in boldface. For C-RO and C-DRO, two different values of l are used.

Table A4 Average MSE and MAE Results and Eliminated Poisoned Samples (EPS) by Data Cleaning Scheme in our Proposed Models under Nopt Attack.

		Nopt											
		0.05			0.10			0.15			0.20		
		MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS	MSE	MAE	EPS
abalone	OLS (no attack)	0.02	0.09		0.02	0.09		0.02	0.09		0.02	0.09	
	LAD (no attack)	0.02	0.09		0.02	0.09		0.02	0.09		0.02	0.09	
	OLS	1.67	0.93		2.64	1.15		3.76	1.32		5.03	1.54	
	LAD	0.06	0.15		0.15	0.24		0.21	0.30		0.22	0.31	
	SR	0.07	0.16		0.15	0.24		0.21	0.29		0.22	0.31	
	DRR	0.02	0.10		**0.04**	0.14		0.05	0.16		0.07	0.19	
	C-RO (l=1-% pois.)	**0.02**	**0.09**	**82%**	**0.05**	**0.12**	**83%**	**0.03**	**0.10**	**82%**	**0.07**	**0.14**	**77%**
	C-DRO (l=1-% pois.)	**0.02**	**0.09**	**89%**	**0.02**	**0.09**	**90%**	**0.02**	**0.09**	**87%**	**0.02**	**0.09**	**84%**
autompg	OLS (no attack)	0.01	0.09		0.01	0.09		0.01	0.09		0.01	0.09	
	LAD (no attack)	0.01	0.10		0.01	0.10		0.01	0.10		0.01	0.10	
	OLS	0.31	0.44		0.43	0.53		0.87	0.79		0.86	0.78	
	LAD	0.04	0.14		0.06	0.19		0.17	0.32		0.22	0.36	
	SR	0.04	0.14		0.10	0.25		0.17	0.32		0.22	0.36	
	DRR	0.04	0.14		0.06	0.19		0.15	0.30		0.17	0.32	
	C-RO (l=1-% pois.)	**0.02**	**0.11**	**53%**	**0.02**	**0.11**	**69%**	**0.02**	**0.11**	**71%**	**0.02**	**0.11**	**70%**
	C-DRO (l=1-% pois.)	**0.02**	**0.11**	**51%**	**0.02**	**0.11**	**67%**	**0.02**	**0.11**	**69%**	**0.02**	**0.12**	**70%**
wine	OLS (no attack)	0.03	0.12		0.03	0.12		0.03	0.12		0.03	0.12	
	LAD (no attack)	0.03	0.12		0.03	0.12		0.03	0.12		0.03	0.12	
	OLS	0.04	0.16		0.10	0.24		0.18	0.32		0.18	0.32	
	LAD	0.03	0.14		0.03	0.14		0.03	0.14		0.04	0.16	
	SR	0.03	0.14		0.03	0.14		0.03	0.14		0.04	0.16	
	DRR	0.03	0.14		0.03	0.14		0.03	0.14		0.04	0.16	
	C-RO (l=1-% pois.)	**0.03**	**0.13**	**42%**	**0.03**	**0.13**	**63%**	**0.03**	**0.13**	**59%**	**0.03**	**0.13**	**65%**
	C-DRO (l=1-% pois.)	**0.03**	**0.13**	**38%**	**0.03**	**0.13**	**58%**	**0.03**	**0.13**	**63%**	**0.03**	**0.13**	**65%**

Detailed results of average of 5 experiments are reported. The best regression model for each case is underlined and marked in boldface, and the second-best model is just marked in boldface. For C-RO and C-DRO, only $l = 1 - \%pois.$ is used since their performance is very close to baseline.

References

Chen R, Paschalidis IC (2018) A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research* 19(13):1–48.

Chen R, Paschalidis IC (2020) Distributionally robust learning. *Foundations and Trends in Optimization* 4(1-2):1–243.

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1):115–166.

Shapiro A (2001) On duality theory of conic linear problems. Goberna MA, López MA, eds., *Semi-Infinite Programming: Recent Advances*, volume 57 of *Nonconvex Optimization and Its Applications*, 135–165 (Boston, MA: Springer).