Approaching Outliers in Dataset with Stable Regression and Minimization of CVaR (Superquantile) Error

AMS518 – Project Report
Zhuoqiao Ouyang
12/1/2024

1.1 Introduction

Linear regression model, often given in the following general form,

$$Y = \beta X + w + \epsilon, (Y, w, \beta) \in \mathbb{R}^n, X \in \mathbb{R}^{n \cdot m}, \epsilon_i \; i.i.d \, N(0, \sigma^2) \quad (1)$$

is widely utilized across different areas of studies as a statistical tool to observe and predict the relationship between independent and dependent variables. The objective when fitting a linear model is often to find the optimal coefficients (w^*, β^*) that minimize the difference or error between the real observed value y_i and the predicted value $y_i = \beta^* X_i + w^*$, denoted by $Z_i(w^*, \beta^*)$, for all observations i.

Under the different contexts of the dataset, additional constraints and objectives might be necessary, which turns the problem of fitting linear models into a linear optimization problem. Also, when fitting a linear model for predicting purposes, one must consider its performance with unobserved data, i.e., the objective is not just to minimize the errors for dataset used for training (in-sample data) but also for inaccessible future data (out-of-sample data). Furthermore, abnormal data points should be identified, and their indication should be analyzed as they have huge impacts on the performance and goodness of fit of the model.

In this project we will discuss the Stable Regression (SR), an optimization approach in training linear model such that the model has optimal out-of-sample performance, with the focus on the underlying problem of fitting SR model to dataset with outliers and the potential of detecting outliers in dataset by fitting the SR model. During the numerical experiments it was also found that outliers may be detected by fitting the dataset to a linear model with the objective of minimizing CVaR (superquantile) error, the error element in the quantile-based quadrangle. The outlier detecting abilities of both models will be compared.

1.2 Background: Cross Validation and Stable Regression

When training a linear regression model, the process of cross validation – where dataset is partitioned into three parts, with one part set as training data (in-sample data), one as validation set and one fixed as reporting set for reporting final result – is often implemented to obtain optimal performance for final reporting in terms of sum of errors $\sum_{i=1}^{n} \left| y_i - \beta^* X_i^T - w^* \right|$ in l_1 space or $\sum_{i=1}^{n} \left| y_i - \beta^* X_i^T - w^* \right|^2$ in l_2 space. Some commonly used partitioning procedures are randomly partitioning the observations into $\alpha\%$ being training set and $(1-\alpha)\%$ being validation

set and the *K*-fold validation procedure – where the observations are first partitioned into *K* parts, then randomly choosing each $K-1$ part as training set and the rest as validation set. As the cross-validation procedures involve randomization, the problem regarding optimally and efficiently choosing a partition of training and validation sets that gives the best out-of-sample performance naturally arises.

In the paper *Stable Regression: On the Power of Optimization over Randomization in Training Regression Problems*, Bertsimas et al. introduced the methodology combining the process of choosing the training set and training the linear model into one linear optimization problem called Stable Regression (in the rest of this report, this optimization problem will be referred to as the SR or the SR model); furthermore under the constraints and objectives of the optimization problem, the training set chosen is the hardest among all, and the model is more stable in the sense that the built models are robust to all subpopulations in data. More specifically, the SR minimization problem has the following objective and constraint:

$$\min_{\beta} \max_{z \in Z} \left(\sum_{i=1}^{n} z_i \left| y_i - x_i^T \beta \right| + \lambda \sum_{i=1}^{p} \Gamma(\beta) \right) \text{ with } Z := \left\{ z_i : \sum_{i=1}^{n} z_i = k,\ z_i \in \{0,1\} \right\} \quad (2)$$

Note that here if $z_i = 1$, then the term $z_i \left| y_i - x_i^T \beta \right|$ is nonzero. The part $\max_{z \in Z} \left(\sum_{i=1}^{n} z_i \left| y_i - x_i^T \beta \right| \right)$ is equivalent to choosing the points $\{(x_i, y_i) s.t. z_i \neq 0\}$ such that the sum of error $\sum_{i=1}^{n} \left| y_i - x_i^T \beta \right|$ in l_1 is maximized with respect to the choice of z_i; then the sum of error is minimized with respect to the choice of (β, λ) by solving the minimization problem $\min_{\beta} \left(\max_{z \in Z} \left(\sum_{i=1}^{n} z_i \left| y_i - x_i^T \beta \right| + \lambda \sum_{i=1}^{p} \Gamma(\beta) \right) \right)$, where $\lambda \sum_{i=1}^{p} \Gamma(\beta)$ is the process of regularization that is usually applied in practice to prevent overfitting and $\Gamma(\cdot)$ is the chosen regularization function. In general, solving (2) is equivalent to optimally, instead of randomly, choosing the hardest set of points (x_i, y_i) as training set of size k in the sense that it leads to maximal sum of errors; then choosing (β, λ) that minimizes such a sum of errors could be understood as training the linear model under the "worst situation", which prepares the model to perform optimally under any other situations.

Given that the training set chosen by SR is the hardest among all subsets, we are concerned with the goodness of fit of the SR model when the dataset contains outliers. Consider an example of a dataset containing 10 observations with one dependent, one independent variable and two outliers. Suppose 20% of the observations are for training, then based on the objective of SR, the chosen training set will include the outliers and cause the model to be biased. Illustration (1) helps visualize this example, where the blue dots represent the fitted SR to the dataset in black dots and clearly the fitted model is tilted towards the outliers. In this case, SR gives neither optimal in-sample nor out-of-sample performances.

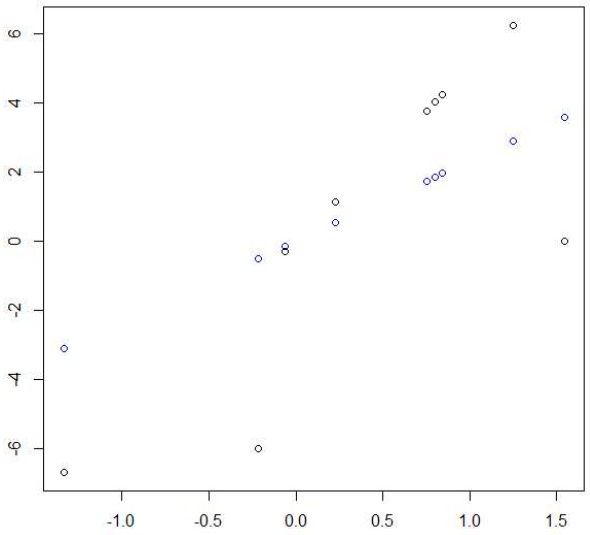

Illustration (1): example of a biased SR model due to outliers in training set

Not only that fitting SR models to datasets with outliers could be troublesome, but in general outliers in datasets should be carefully identified and processed before applied to training any kind of model to prevent bias and maintain accuracy. Motivated by the above example and the property of training set selected by SR, since outliers in dataset also lead to large sum of errors, we are interested in testing whether fitting SR can be utilized as a technique for identifying and removing the outliers. More specifically, suppose SR model is fitted for a dataset of n observations with $\alpha\%$ of outliers and $\alpha\%$ of the observations are set as training set, we check if the optimal z^* to (2) is such that $\sum_{i=1}^{n} z_i^* = \alpha\% \cdot n$ and the corresponding (x_i, y_i) chosen for the training set are the outliers.

2.1 Algorithm for Solving the SR Optimization Problem and Identifying the Underlying Training Set

The algorithm for solving the SR minimization problem provided in Bertsimas et al involves the techniques of introducing dual variables. In this project, instead of solving the SR, we will be solving its equivalent ν - SVR (Nu Support Vector Regression), which was introduced in *Support Vector Regression (SVR): Risk Quadrangle Framework* [Malandii and Uryasev, 2023]. Similar to the SR, the ν -SVR is a linear optimization problem.

First consider the following notations and definitions:

1. Risk Quadrangle, introduced in *The Fundamental Risk Quadrangle in Risk Management, Optimization and Statistical Estimation* [Rockafellar and Uryasev, 2013], links risk measurement, reliability, statistics, and stochastic optimization in one framework. Each quadrangle contains five elements: risk, deviation, error, regret, and statistics. Some examples of the risk quadrangle include the quantile-based quadrangle and CVaR (superquantile) norm quadrangle.

2. Let $Z(w,\beta) = (y_1 - x_1^T\beta - w, \ldots, y_n - x_n^T\beta - w)$ be the loss vector, where (w,β) is the pair of coefficients to the linear model in (1) and (x_i, y_i) are the observations of the independent and dependent variables X and Y, respectively.

3. Let X be a real valued random variable, then $q_\alpha(X)$ denotes the VaR (value at risk) with parameter $\alpha \in [0,1]$ and $\overline{q}_\alpha(X)$ denotes the CVaR (conditional value at risk) of X with parameter $\alpha \in [0,1]$. $q_\alpha(X)$ is the statistic and $\overline{q}_\alpha(X)$ is the risk measurement in the quantile-based quadrangle.

4. Let $\langle\langle X\rangle\rangle_\alpha$ denotes the CVaR norm, which is the error measurement in the CVaR norm quadrangle. The scaled CVaR norm is defined as $\langle\langle X\rangle\rangle_\alpha = \overline{q}_\alpha(|X|)$ and the non-scaled CVaR norm is defined by $\langle\langle X\rangle\rangle_\alpha = (1-\alpha)\overline{q}_\alpha(|X|)$.

With reference to these notations and definitions, the ν-SVR is linear regression minimization problem having the following general formulation:

$$\min_{w,\beta} \langle\langle Z(w,\beta)\rangle\rangle_\alpha + \frac{\lambda}{2}||w|| \quad (3)$$

It can be shown [Malandii and Uryasev, 2023] that (3) can be equivalently rewritten as the following form:

$$\min_{w,\beta} \max_{q \in Q_\alpha} \quad \sum_{i=1}^n q_i |z_i(w,\beta)| + \lambda||w||_2^2 \text{ with } Q_\alpha := \left\{q \in \mathbb{R}^n : \sum_{i=1}^n q_i = 1, 0 \le q_i \le \frac{1}{n(1-\alpha)}\right\} (4)$$

Note that by letting $k = n(1-\alpha)$ and $z_i = k \cdot q_i$, $\sum_{i=1}^n z_i = k$ and $0 \le z_i \le 1$, which proves that $Q_\alpha = Z := \{z_i : \sum_{i=1}^n z_i = k,\ 0 \le z_i \le 1\}$, the convex hull of Z in (2). Since it can be proved [Bertsimas et al, 2020] that solving (2) with Z is equivalent to solving (2) with the convex hull of Z, one may conclude that the SR problem of (2) is equivalent to the ν-SVR problem in (4). Given the equivalence of the two problems, for the rest of this report, ν-SVR will be referred to as SVR, and SR and SVR will be treated as the same model.

Furthermore, it can be shown [Malandii and Uryasev, 2023] that given optimal solution (w^*, β^*, q^*) to (4), the nonzero optimal q_i^* are the ones where $|z_i(w^*,\beta^*)| \ge q_\alpha(|Z(w^*,\beta^*)|)$, hence the training set chosen by SVR consists of points (x_i, y_i) such that the corresponding $z_i(w^*,\beta^*) = (y_i - x_i^T\beta^* - w^*)$ satisfies the above inequality. Overall, suppose a dataset with n observations is given and that $(1-\alpha)\%$ of the observations are to be used for training, the detailed procedures for solving SVR and identifying the underlying training set are as follows:

1. Solving SVR with parameter α to obtain the optimal coefficients (w^*, β^*); this can be done by applying the $(1-\alpha)cvar\_risk(\alpha, abs(\cdot))$ function from PSG (Portfolio Safeguard) to the scenario matrix X.

2. Calculate $Z(w^*, \beta^*)$ and $q_\alpha(|Z(w^*, \beta^*)|)$ with the PSG $var\_risk(\cdot)$ function.

3. Compare each $|z_i(w^*, \beta^*)|$ with $q_\alpha(|Z(w^*, \beta^*)|)$: if $|z_i(w^*, \beta^*)| \geq q_\alpha(|Z(w^*, \beta^*)|)$, then the corresponding points (x_i, y_i) are selected as the training set.

4. In the case of testing whether the SVR model can identify outliers, randomly choose $(1-\alpha)\%$ of the observations to be outliers and repeat step 1-4. Compare the index i where $|z_i| \geq q_\alpha(|Z(w^*, \beta^*)|)$ with the index j of the outliers in the observations. If $i = j$ for some j, then we say that the SVR model has correctly identified an outlier y_i. We define the outlier detecting ability by the percentage $ODA := \frac{number\ of\ outliers\ correctly\ identified\ by\ SVR}{total\ number\ of\ outliers}$

2.2 Detecting Outliers by Fitting Linear Model that Minimizes CVaR Error.

During the study it was also noticed that solving linear optimization problem that minimizes the CVaR error (in the rest of the report, this optimization problem will be referred to as the CVaR error model), which is the error element in the quantile-based quadrangle, together with a series of similar procedures presented in section 2.1, might enable one to identify the outliers in the dataset. To test such capability for the CVaR error model, first suppose that a given dataset has $(1-\alpha)\%$ of the observations being outliers, then follow the below procedures:

1. Solving CVaR error minimization problem with parameter α to obtain optimal (w^*, β^*): this can be done by applying the PSG function $cvar2_err(\cdot)$with parameter α to the scenario matrix X. Calculate $Z(w^*, \beta^*)$ and $q_\alpha(|Z(w^*, \beta^*)|)$.

2. Compare each $|z_i(w^*, \beta^*)|$ with $q_\alpha(|Z(w^*, \beta^*)|)$. Compare each index i where $|z_i(w^*, \beta^*)| \geq q_\alpha(|Z(w^*, \beta^*)|)$ with the index j of the outliers in the observations. If $i = j$ for some j, then we say that the CVaR error model has correctly identified an outlier y_i. We define the outlier detecting ability by the percentage

$$ODA := \frac{number\ of\ outliers\ correctly\ identified\ by\ CVaR\ error\ model}{total\ number\ of\ outliers}$$

Section 3. Case Study

This section will present the case study in solving the SVR and CVaR error minimization problems and testing whether both models can identify outliers under the theoretical procedures described in 2.1 and 2.2 using real numerical data. For simplicity, the SVR problem in form (4) will be solved with regularization term dropped. The dataset used will be synthetic generated data with only one independent variable and no intercept, alternated such that 5% are outliers. The goal is to detect all 5% outliers, thus the size of training set in SVR is chosen to be $5\% \cdot n$, where n is the number of observations, and the confidence level α is chosen to be 0.95;

similarly, the confidence level for solving the CVaR error minimization problem is chosen to be 0.95. For the result, we report the ODA of both models.

3.1 Design of Synthetic Data and Results

The datasets used are randomly generated linear models, where independent variable X is a vector of length n randomly generated from $N(0,1)$ (normal distribution mean 0 and variance 1), noise components ϵ with $\epsilon_i \sim N(0,1)$, unobserved true regressionβ_{true} is a vector of length n generated from uniform distribution with a lower bound randomly generated from $U(-200,200)$and an upper bound equal to $lower\ bound \cdot 1.5$; lastly the dependent variable Y is defined by $y_i = \beta_{true\,i} x_i + \epsilon_i$. The dataset is then alternated such that 5% of the observations are outliers in the following way: first randomly draw 5% of the observations y_i, then each y_i is multiplied with a number randomly generated from $U(9,10)$ and add to a number randomly chosen from $\{\max(Y) \cdot a, \min(Y) \cdot b\ with\ a \in U(-2,2), b \in (-3,-1)\}$. Now the scenario matrix is one with columns of scenario benchmarks adjusted with 5% being outliers, noise components and X . Here's an example of the generated dataset:

	Y_syn	rand_num_noise	rand_num_X	Outlier_Position
[1,]	118.67945	-0.2429203	2.4405896	22
[2,]	-72.40861	1.0041214	-1.6324859	46
[3,]	-32.38144	-1.3318255	-0.6922465	50
[4,]	80.46493	-1.7084688	1.6016069	0
[5,]	42.34127	-0.6754959	0.8851593	0
[6,]	24.69100	0.2061206	0.4942090	0

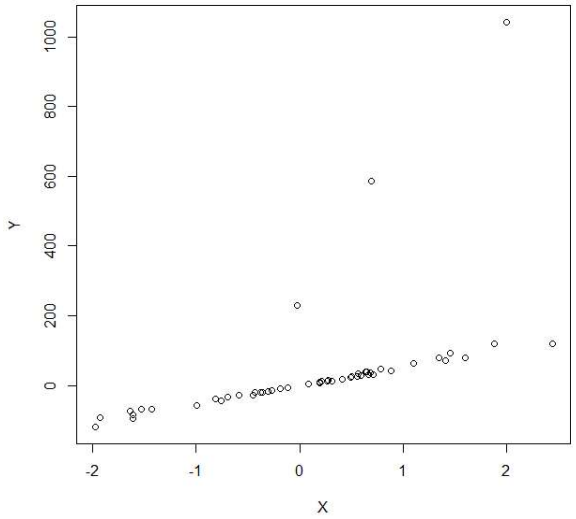

Illustration (2): example of generated dataset

Synthetic datasets with observations number $n = 50,100,500,1000$ are used for the case study; for each n , the experiment was repeated 10 times, and the ODA of both SVR and CVaR error models are reported for each experiment and their averages are reported.

n = 50	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**	**AVG**
SVR O.D.A	0.667	0.667	0.667	0.667	0.667	1	0.667	1	0.667	0.667	0.7336
CVaR Error O.D.A	1	1	1	1	1	1	0.667	1	0.667	1	0.9334
n = 100	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**	**AVG**
SVR O.D.A	0.833	0.667	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.833	0.8164
CVaR Error O.D.A	0.833	0.833	1	1	1	1	1	1	1	0.833	0.9499
n = 500	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**	**AVG**
SVR O.D.A	0.962	0.885	0.808	0.885	0.846	0.885	0.962	0.962	0.885	0.846	0.8926
CVaR Error O.D.A	0.962	0.885	0.962	1	0.846	0.885	0.962	1	1	0.846	0.9348
n = 1000	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**	**AVG**
SVR O.D.A	0.902	0.922	0.882	0.863	0.843	0.863	0.843	0.902	0.882	0.902	0.8804
CVaR Error O.D.A	0.902	0.922	1	1	0.98	0.98	0.843	0.902	0.941	0.902	0.9372

3.2 A Discussion of the Results

Based on the tables, it can be observed that the ODA of the SVR model increases and is more stable as the number of observations in the dataset increases: when $n = 50$, the range of the ODA is 0.333, and when $n = 1000$, the range is 0.079; overall, the SVR model is at least 80% correct in detecting the abnormal data points when $n > 100$ and this percentage is close to 90% when n is large. In comparison, the average ODA of the CVaR error model maintains constant between 93% and 95% for all n, and as n increases, the ODA becomes more stable for each numerical experiment (the range of ODA is 0.157 for $n = 1000$ in comparison to the range being 0.333 when $n = 50$). Overall, the CVaR error model has a higher accuracy than the SVR model in detecting outliers.

Scatter plots were drawn to help visualize the results. For $n = 50$, we observe that fitted SVR model is more likely to be impacted by the outliers in the sense that the slope of the fitted line is closer to the slope of the fitted line of the abnormal points, which cause the fitted values to deviate from the dataset. But also, since the SVR model is objected to minimize the errors between such fitted line and the abnormal points, when an abnormal point lies close to the fitted line, the point will not be detected. Illustration (3) *– the visualization of test 2 – is an example

of such a situation. On the other hand, the fitted values of CVaR error model are more closely aligned with the dataset, and hence the model is more sensitive to any abnormal points. The case where the ODA for CVaR error model is low is often since an undetected abnormal point lies within the set of normal points, where in this case, the "abnormal point" cannot be defined as an outlier. Illustration (4) *– the visualization of test 7 – is an example of such a situation. (*Refer to the appendix for illustrations).

Both potential causes for ODA to be less than one for both models remain as n increases. When n is large, we observe that the proportion of the abnormal points lying close to the fitted SVR model decreases, which causes fewer negative impacts on the ODA and hence ODA for SVR model increases. It can also be noticed that the proportion of the abnormal points lying within the normal dataset decreases but the possibility increases, hence the ODA of CVaR error model remains between 90% and 100% in almost all cases.

To conclude, the SVR and CVaR error models enable one to detect outliers in a dataset with a certain degree of inaccuracies that decreases gradually as the number of observations increases. Such inaccuracies are most likely dependent on the position of the designed outliers – if all abnormal points are further away from the original normal dataset, then we expect the detecting abilities for both models to be more accurate. Furthermore, when an abnormal point is close to or lies within the set of normal data points, failure to remove the abnormal point does not necessarily lead to a worse sum of errors. Numerical experiments with the following steps were performed to support this claim:

1. Remove outliers detected by the SVR model and CVaR error model from the original dataset, then fit a linear model that minimizes the mean absolute error with PSG function $meanabs\_err(\cdot)$ and obtain the mean absolute error; denote these mean absolute errors by mae1 and mae2, respectively.

2. Remove the actual outliers from the original dataset and fit a linear model that minimizes the mean absolute error and obtain the mean absolute error; denote this mean absolute error by mae3.

Steps 1 and 2 were repeated 10 times for each $n = 50,100,500,1000$; mae1, mae2, mae3, ODA of both models, and the percentage in which mae1 exceeds mae2 are reported:

n = 50	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**
MAE1	37.435	84.492	0.736	20.924	62.983	10.424	41.120	19.047	6.959	41.472
MAE2	16.523	84.492	0.736	4.991	62.983	10.424	15.643	19.047	6.959	9.994
MAE3	16.523	84.492	0.736	4.991	62.983	10.758	15.643	19.047	6.959	9.994
SVR O.D.A	0.667	1	1	0.667	1	0.667	0.667	1	1	0.667

CVaR Error O.D.A	1	1	1	1	1	0.667	1	1	1	1
Difference	126.56%	0.00%	0.00%	319.20%	0.00%	-3.10%	162.87%	0.00%	0.00%	314.99%

n = 100	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**
MAE1	11.606	24.387	22.070	55.905	17.237	68.385	55.158	13.390	45.505	91.080
MAE2	6.357	19.609	8.749	54.855	17.237	68.606	55.158	11.556	45.196	85.321
MAE3	6.357	19.609	8.749	57.484	17.226	68.686	56.734	11.556	45.196	85.321
SVR O.D.A	0.833	0.833	0.667	0.833	0.833	0.833	0.833	0.833	0.833	0.833
CVaR Error O.D.A	1	1	1	0.833	0.833	0.833	0.833	1	1	1
Difference	82.56%	24.37%	152.26%	-2.75%	0.06%	-0.44%	-2.78%	15.88%	0.68%	6.75%

n = 500	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**
MAE1	53.168	59.031	56.471	16.927	20.681	79.494	90.405	6.545	2.612	85.183
MAE2	49.430	59.031	56.365	13.028	16.318	79.494	90.166	6.302	2.604	85.228
MAE3	49.833	59.381	57.071	13.028	16.318	79.727	90.836	6.367	2.604	87.470
SVR O.D.A	0.846	0.923	0.885	0.808	0.846	0.962	0.885	0.923	0.962	0.846
CVaR Error O.D.A	0.962	0.923	0.885	1	1	0.962	0.885	0.962	1	0.846
Difference	6.69%	-0.59%	-1.05%	29.93%	26.74%	-0.29%	-0.47%	2.80%	0.29%	-2.61%

n = 1000	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test 10**
MAE1	92.479	2.644	10.333	6.449	23.364	5.231	1.810	73.224	29.390	26.343
MAE2	92.031	2.524	8.501	5.703	23.386	5.181	1.760	73.006	29.368	26.343
MAE3	94.316	2.549	8.526	5.753	23.658	5.197	1.768	74.996	29.921	26.861
SVR O.D.A	0.882	0.882	0.841	0.863	0.941	0.922	0.902	0.863	0.902	0.882
CVaR Error O.D.A	0.882	0.961	0.961	0.961	0.941	0.961	0.961	0.863	0.902	0.882
Difference	-1.95%	3.74%	21.19%	12.11%	-1.24%	0.66%	2.39%	-2.36%	-1.77%	-1.93%

The table demonstrates that the mean absolute errors do not always differ from each other even in the case when the SVR and CVaR error models fail to remove all actual outliers, which shows that the real accuracy of both models in detecting outliers is higher than as shown by the numerical values of the ODA.

Section 4. Application of Outliers Detection Techniques and Conclusion

The strength of fitting a SR model or equivalently a SVR model is that it prevents the randomization process in partitioning and selecting the training data set; however when the dataset contains outliers and the size of the training set is small, there exists the risk of training the model based on outliers and leading to biased in-sample and out-of-sample performance. The results presented in section 3 show that this property of the SVR, on the other hand, can be used to detect outliers in the dataset, and the outlier detecting ability is at least 80% accurate when $n > 50$.

While in the reality one cannot always foresee how many outliers there will be in a dataset – especially in unobserved datasets – and the definition of outliers vary under different contexts, if certain proportion of the in-sample dataset can be discarded, then the SVR model and CVaR error model can be used to remove the abnormal points that could potentially lead to bias. Numerical experiments were performed to demonstrate the effect of removing outliers on out-of-sample performances, where in the experiments 30% of the synthetic generated dataset (with 10% outliers) are fixed as out-of-sample data, then under the assumption that 5% of the data can be discarded, SVR model and the CVaR error model were used to detect and remove the 5% "hardest" data points. SVR model was fitted to the processed datasets with $k = 10\%, 20\%, 40\%, 60\%, 80\%$ being the desired proportions of training data and out-of-sample mean absolute errors were obtained. These mean absolute errors were then compared to the out-of-sample mean absolute errors obtained from the SVR model fitted to the unprocessed datasets with $k = 10\%, 20\%, 40\%, 60\%, 80\%$ being the desired proportions of training data. Below are the results for the dataset with $n = 50,100,500,1000$ observations, and mae1, mae2, mae3 denote the out-of-sample mean absolute errors for the three cases, respectively.

n = 50	**k=10%**	**k=20%**	**k=40%**	**k=60%**	**k=80%**
MAE1	817.610	306.703	324.662	321.066	315.453
MAE2	322.376	301.407	324.662	322.667	331.307
MAE3	1005.72	325.89	329.76	334.81	332.43
n = 100	**k=10%**	**k=20%**	**k=40%**	**k=60%**	**k=80%**
MAE1	222.850	222.836	222.814	222.803	222.908
MAE2	222.960	222.847	222.717	222.712	222.786
MAE3	675.918	222.751	222.798	222.755	222.716
n = 500	**k=10%**	**k=20%**	**k=40%**	**k=60%**	**k=80%**
MAE1	418.280	417.831	417.787	417.739	417.781
MAE2	417.856	417.792	417.764	417.742	417.798
MAE3	552.875	418.462	417.751	417.765	417.770
n = 1000	**k=10%**	**k=20%**	**k=40%**	**k=60%**	**k=80%**

MAE1	175.514	175.322	175.266	175.296	175.211
MAE2	175.159	175.278	175.222	175.261	175.165
MAE3	441.666	175.554	175.554	175.524	175.330

These results show that when the size of the training set is small, identifying and removing 5% of the outliers before fitting SVR, regardless of accuracy, can effectively improve out-of-sample performance for the SVR model. As the size of the training set increases, since more normal points are selected to be within the training set, such effect is less significant.

It can be noticed that the optimal out-of-sample performances (highlighted in blue) are often produced by fitting the SVR model after removing outliers using the CVaR error model – though the values are relatively close to the results produced by removing outliers with SVR model. This observation matches the results from section 3.2 since the CVaR error model has the highest accuracy in detecting outliers.

Overall, we conclude that fitting the SVR model is an effective and reliable technique in removing outliers in dataset with large number of observations ($n > 100$), and removing outliers before fitting the SVR model is suggested when training set size is small ($< 20\%$ of total observations); fitting the CVaR error minimization model could be another potential method with higher accuracy in detecting outliers, however such capability require more theoretical support, and that the CVaR error model is less robust than SVR model in the sense such that it cannot determine the optimal dataset for training that gives best out-of-sample performance.

A. Appendix

A.1 References

D. Bertsimas and I. Paskov. Stable regression: On the power of optimization over randomization. Journal of Machine Learning Research, 21(230):1–25, 2020.
URL http://jmlr.org/papers/v21/19-408.html.

Malandii, A. and S. Uryasev. Support Vector Regression: Risk Quadrangle Framework, Jan 2023, arXiv:2212.09178

R. T. Rockafellar and S. Uryasev. The Fundamental Risk Quadrangle in Risk Management, Optimization and Statistical Estimation. Surveys in Operations Research and Management Science, 18(1):33—53, 2013

A.2 Graphs

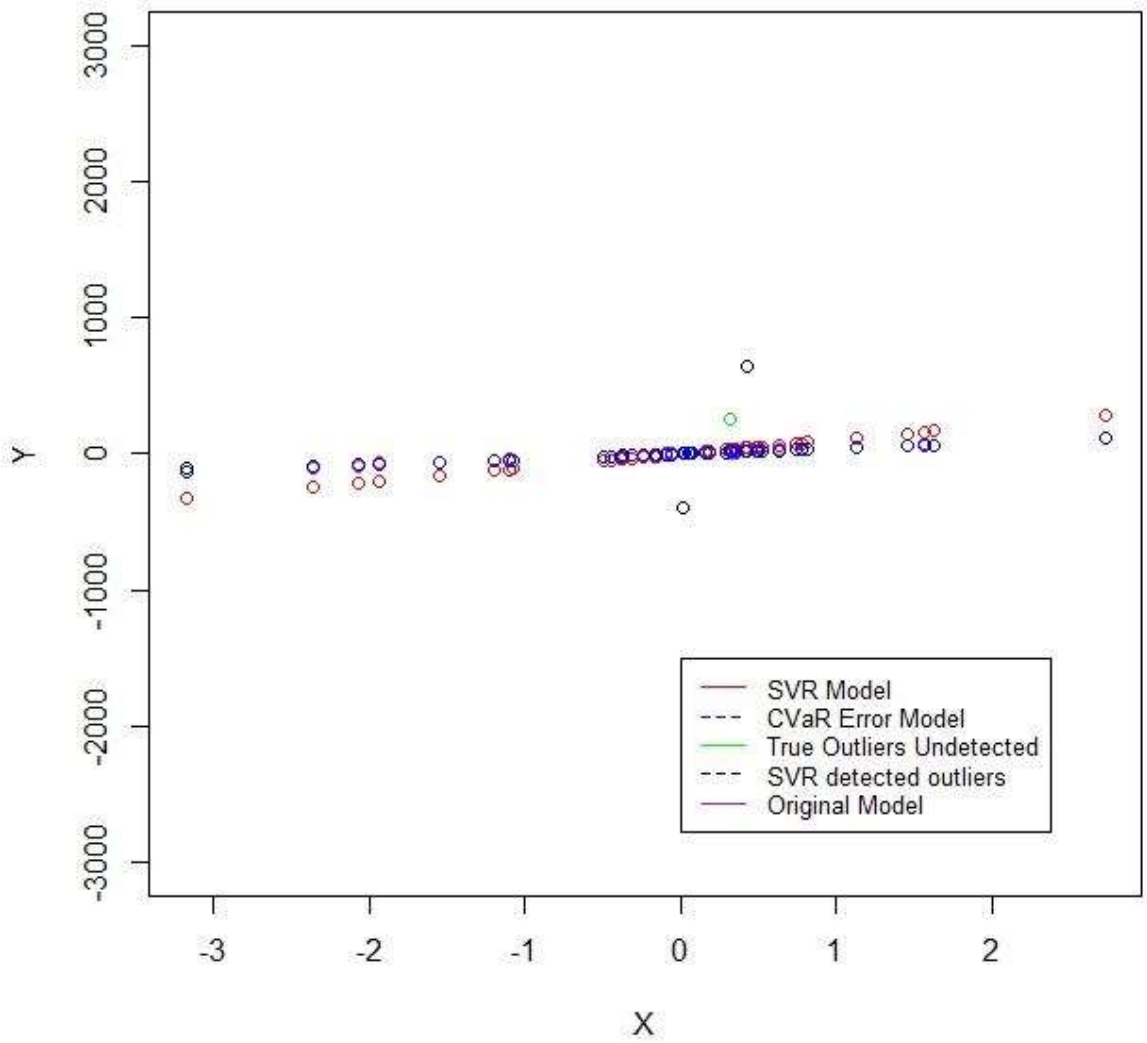

Illustration (3): $n = 50\ test\ 2$

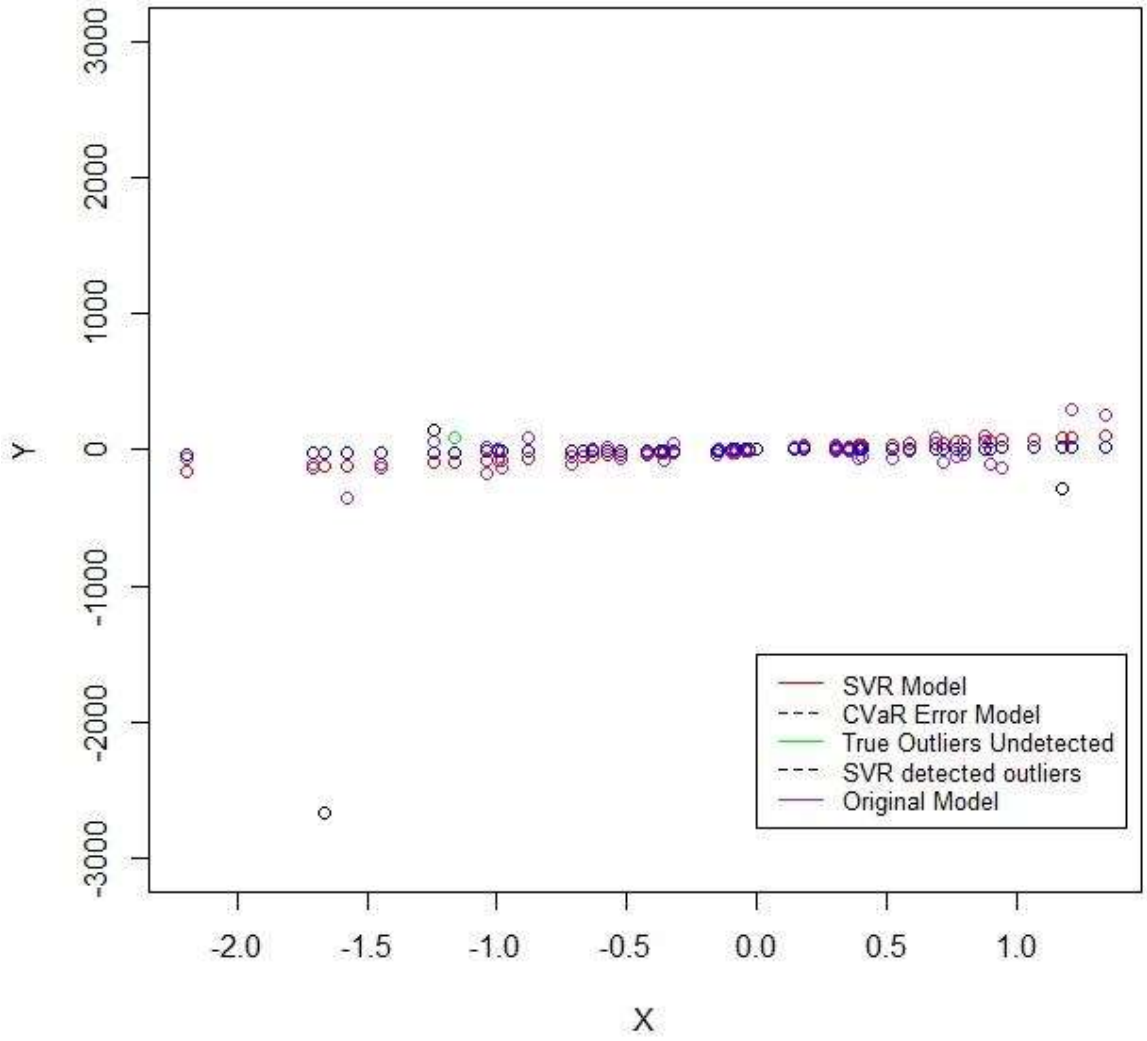

Illustration (4): $n = 50$ *test* 7

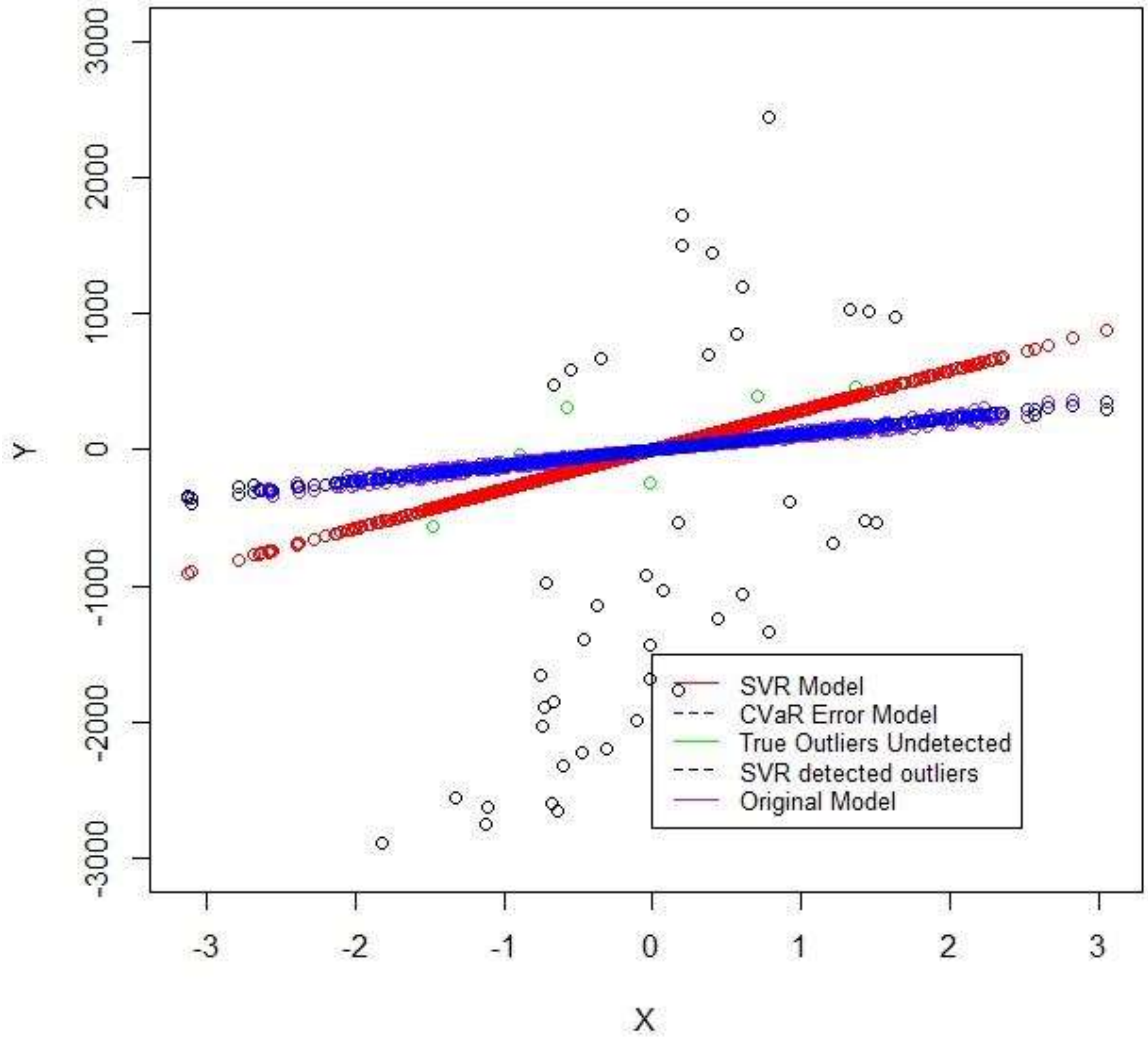

Illustration (5): $n = 1000\ test\ 3$

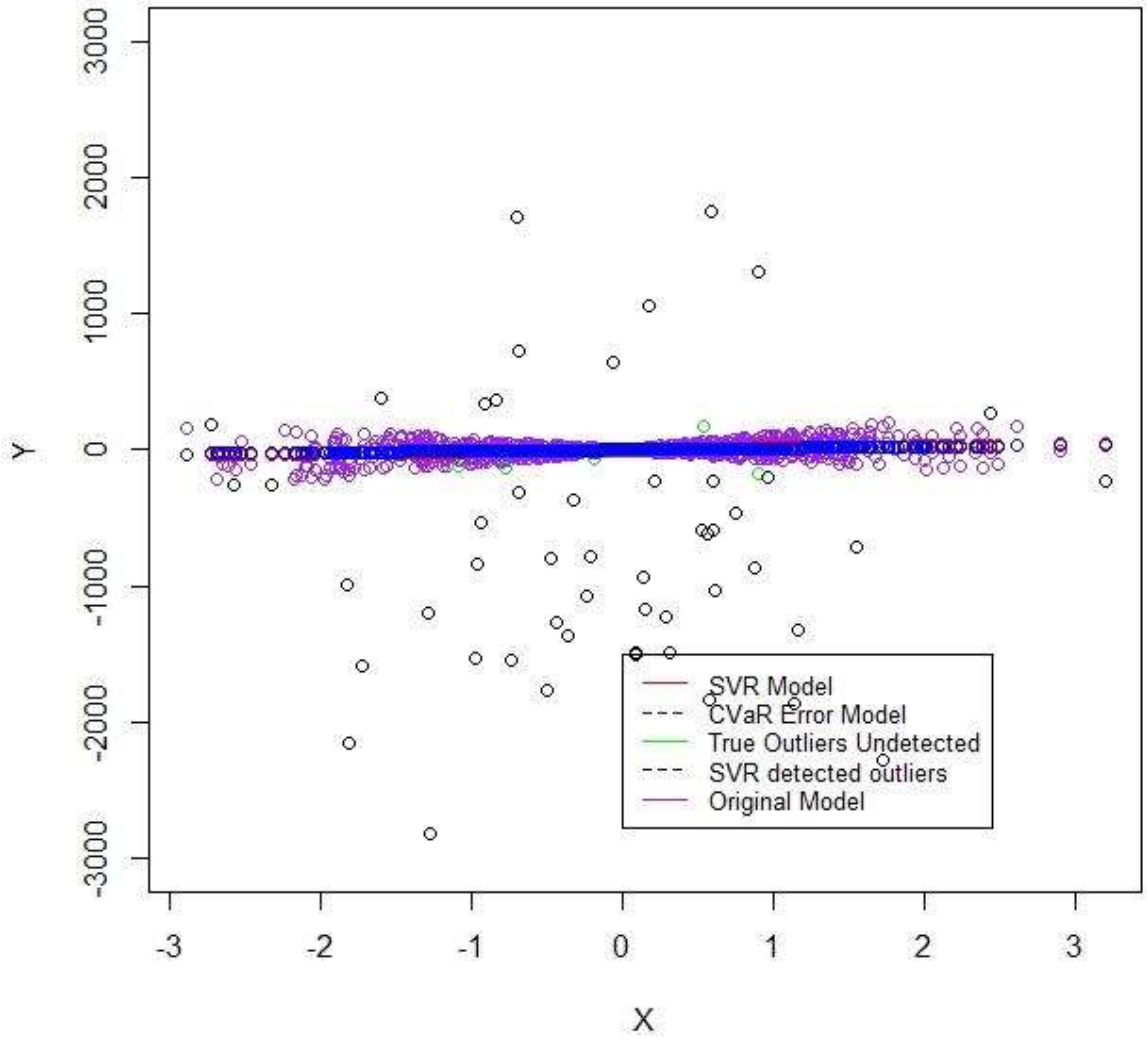

Illustration (6): $n = 1000\ test\ 8$