# Remedies for Outlying Observations in Survival Analysis

By Ente Kang

*In Survival Analysis, our outcome of interest is the time it takes for an event to happen. The validity of estimations is then put into question with the presence of outliers. Using data from patients receiving pembrolizumab, the removal of outlying observations increased the predictive performance of the Cox Proportional Hazards model by 9%. This implies that model estimates should be interpreted with caution and emphasizes the importance of model diagnostics.*

## 1. Introduction

Statistical models are deployed in many areas such as public health and medicine. When attempting to unravel mysteries, researchers must pay close attention to the effect of outlying observations in their studies. Does it really make sense to remove data points just because it is unusual, and will our models still behave nicely if these points are kept in?

The widely used Cox Proportional Hazards model for Survival Analysis suffers with the presence of outlying points. It has a breakdown point of $\frac{1}{n}$, meaning that just one outlying observation will introduce biases into the estimation. In this paper, we will introduce outlier detection algorithms, and extensions that have been made to the Cox model to make it more robust.

In this section, we will introduce the basics of the Cox model. Section 2 contains details of algorithms used for detecting outliers and extensions of the Cox model, In section 3, these methods are applied to a real-world dataset, and the results are then discussed. Lastly, section 4 will summarize the key points, address limitations, and provide future recommendations.

### 1.1 Cox Proportional Hazards Model

Let *T* be the event time of interest, *C* be the censoring time, and $\delta = I(T \leq C)$ which is a censoring indicator.

Censoring is when we do not observe the event time $T_i$. For example, a patient may drop out of the study before the study endpoint, or a patient might not experience the event during the study period. In all scenarios, we only have information on the time they were censored ($C_i$), i.e., censoring time.

Let the Hazards function be $h(t|X) = \mathrm{h}_0(t)\exp\{X^T\beta\}$, where t denotes the time, X is a vector of covariates, and $\beta$ a vector of regression coefficients. The hazards function can be interpreted as the intensity of some event happening at time t, given that an individual has not experienced the event for at least t periods. This is a semi-parametric approach because we do not make any assumptions about the function form of $h_0(t)$ - the baseline hazards. This method is not robust against outliers, and has a breakdown point of $\frac{1}{n}$ [1]. Meaning that all it takes is one observation to introduce biases into the estimation.

Fortunately, there are ways to remedy this. We will consider model-based approaches, where we adjust the model specification and methods to identify outliers based on various diagnostic tests.

## 2. Methods:

### 2.1 Robust Cox Regression

A slight modification of the Cox Proportional Hazards model introduces robustness in the estimation. The new regression coefficients $\hat{\beta}$ are determined through solving the following equation.

$$\sum_{\{i=1\}}^{n} w_i\delta_i[x_i - \frac{\sum_{\{t_j\geq t_i\}} w_{ij}x_j\exp\{\beta^T x_j\}}{\sum_{\{t_j\geq t_i\}} w_{ij}\exp\{\beta^T x_j\}}] = 0$$

In the outer summation, $w_i$ downweighs the uncensored observations ($\delta_i = 1$). In the inner summation, $w_{ij}$ downweighs individuals with large values of $\exp\{\beta^T x_j\}$. In other words, we are doubly weighing the partial likelihood function. The weight functions are typically linear, quadratic, or exponential.

To see why this works, note that regression coefficients $\beta$ in the traditional Cox Proportional Hazards model are obtained through the maximization of the following partial likelihood function,

$$\prod_{\{i=1\}}^{n}\left[\frac{x_j\exp\{\beta^T x_j\}}{\sum_{\{t_j\geq t_i\}}\exp\{\beta^T x_j\}}\right]^{\delta_i}$$

Individuals with unusually low or high survival times for a given set of features will bias the estimations of $\beta$.

### 2.2 Censored Quantile Regression

Censored Quantile Regression is a model that has been gaining attention in health and medicinal research.

$$Q_{T_i}(\theta|x_i) = X^T\beta(\theta)$$

Here $\theta \in (0,1)$ are the quantiles, and Q is the quantile function. Notice that the regression coefficients $\beta$ are a function of $\theta$. This allows for the regression coefficients to be different at different quantiles of the time the event occurs, $T_i$.

The subtlety of this method is to be appreciated as it accounts for the heterogeneity of individuals by acknowledging that people at different quantiles of some event of interest will not respond to treatment the same way. For example, individuals with a very long survival time may respond to medication differently than someone with a much shorter survival time.

With $\beta$ varying for different quantiles of the response variable, this method is more robust against outliers since the effect of covariates is no longer global, but rather local. Multiple algorithms exist for estimation [5] of $\beta$ such as Portnoy, Bottai and Zhang or Wang and Wang. The computational details are outside the scope of this paper [5] as our focus is on treatment of outliers in survival analysis.

2.3 Concordance C-index based methods

In this section, the Concordance C-index *(figure 1)* will first be introduced as a precursor to outlier detection algorithms that are based on it.

Let $T_i, T_j$ denote the occurrence of the event for individuals i and j. The censoring times are $C_i$ and $C_j$. Furthermore, let $\lambda$ denote the risk, which can be interpreted as the hazards for a person.

1. Form all permissible pairs by omitting pairs where the shorter event time is censored or cases where both individuals are censored.
2. For each of these permissible pairs, do the following.
    a. If the event time is longer for an individual with lower hazards $\lambda$, count 1; otherwise count 0.5. If both members of a pair have the same event time and are both not censored, with the same hazard, count 1; otherwise count 0.5. If one member is censored, while the other is not, with the censored individual having lower hazards, then count 1; otherwise count 0.5.
3. The concordance is the sum of these counts. Dividing by the number of permissible pairs, N, gives us the c-index.

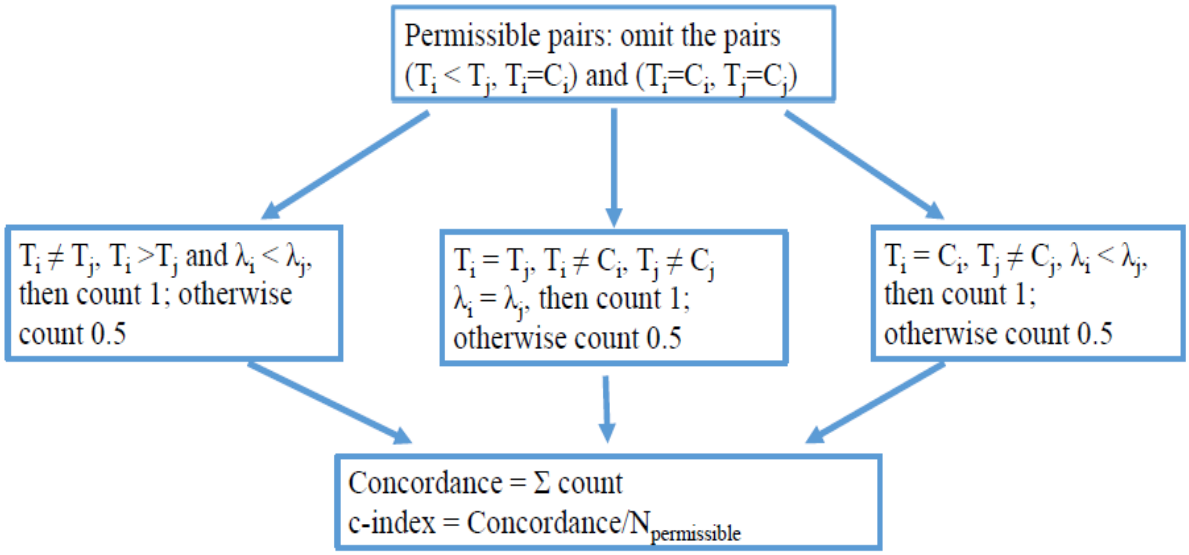


*Figure 1*

In essence, concordance means that the prediction tends towards the same direction as the data suggests. Let $y_i$ and $y_i'$ be the true and estimated values of the response variable for individuals i and j. A pair is considered concordant if $y_i > y_j \; and \; y_i' > y_j'$ or $y_i < y_j \; and \; y_i' < y_j'$. For the Cox model, a longer survival time corresponds to a lower hazard, i.e., smaller $X^T\beta$.

Ideally, we want to see a concordance c-index close to 1. A c-index of 0.5 is no better than random guessing. This metric will be used to assess the performance of our models.

### 2.3.1 One Step Deletion (OSD)

This is the first algorithm that we will explore, and works in a sequential fashion.

1. Using the full dataset, we calculate the concordance c-index by removing the $i^{th}$ observation. For data of size n, there will be n concordance values.
2. The observation when removed that leads to the largest increase in concordance is removed.
3. With our new dataset with n-1 observations, we iterate through steps (1) and (2) until a specified number of outliers are removed.

### 2.3.2 Bootstrap Hypothesis Test (BHT)

Bootstrap Hypothesis Testing is built on the idea of the bootstrap (Efron, 1979), where we repeatedly sample the data at hand *(figure 2).*

Let $C_{-i}$ denote the concordance c-index for the model with the ith observation removed, and $C_{original}$ be the c-index with the full data set. For each one of the data points, our hypothesis is as follows.

$$H_0: C_{-i} \leq C_{original}$$

$$H_A: C_{-i} > C_{original}$$

Which can be reformulated as,

$$H_0: \delta C_i \leq 0$$

$$H_A: \delta C_i > 0$$

Where $\delta C_i = C_{-i} - C_{original}$.

For a dataset of size n, the procedure works as follows.

1. For each observation i, remove it and generate B bootstrap samples from the n – 1 observations.
2. Calculate the concordance for each of these B bootstrap samples
3. The p-value is the fraction of bootstrap samples having $\delta_i \leq 0$
4. Repeat steps (1) – (3) for all observations.

A small p-value for any of the n hypothesis tests is evidence suggesting that the removal of observation i leads to an increase of the c-index. The number of bootstrap samples B, depend on the size of the dataset and the number of variables in the model.

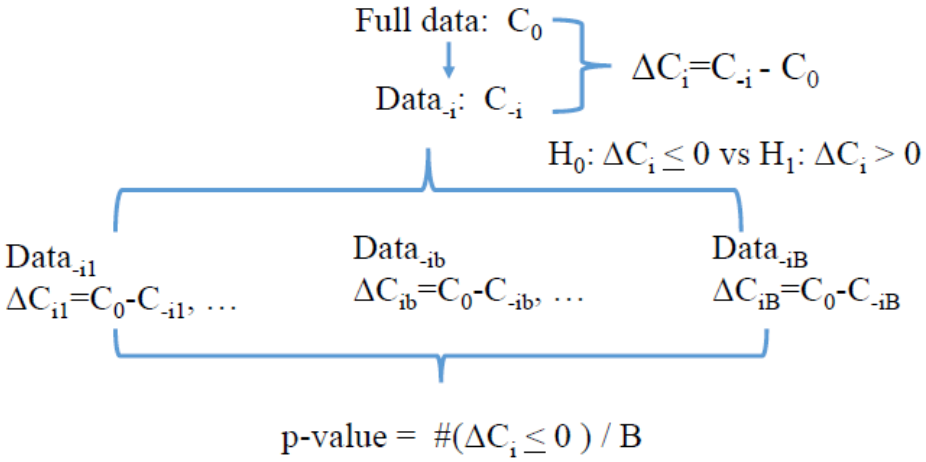


*Figure 2*

In general, caution is needed when performing bootstrapping when the dataset has many observations, as it increases the computational burden.

## 3. Results

The *pembrolizumab* dataset was used to assess these outlier detection methods, which had 73 observations and 8 variables (*table 1*).

| Predictors | Full Sample (n=73) | Censored (n=11) | Non Censored (n=62) | p-value |
|---|---|---|---|---|
| **age** | | | | 0.19 |
| Mean (sd) | 57.9 (13.4) | 63.0 (13.1) | 57.0 (13.3) | |
| Median (Min,Max) | 59.5 (21.1, 81.8) | 67.8 (34.1, 81.8) | 57.9 (21.1, 78.2) | |
| **sex** | | | | 1.00 |
| Female | 40 (55) | 6 (55) | 34 (55) | |
| Male | 33 (45) | 5 (45) | 28 (45) | |
| **cohort** | | | | **0.02** |
| A | 14 (19) | 2 (18) | 12 (19) | |
| B | 11 (15) | 0 (0) | 11 (18) | |
| C | 10 (14) | 0 (0) | 10 (16) | |
| D | 10 (14) | 5 (45) | 5 (8) | |
| E | 28 (38) | 4 (36) | 24 (39) | |
| **l size** | | | | 0.66 |
| Mean (sd) | 85.9 (59.9) | 84.4 (60.2) | 86.2 (60.3) | |
| Median (Min,Max) | 73 (11, 387) | 64 (33, 241) | 74.5 (11.0, 387.0) | |
| **pdl1** | | | | **0.05** |
| Mean (sd) | 15.1 (30.9) | 37.2 (45.3) | 11.2 (26.3) | |
| Median (Min,Max) | 1 (0, 100) | 12 (0, 100) | 0 (0, 100) | |
| **tmb** | | | | 0.10 |
| Mean (sd) | 1.0 (1.1) | 1.7 (1.8) | 0.8 (0.8) | |
| Median (Min,Max) | 0.7 (-0.8, 5.2) | 1.4 (-0.8, 5.2) | 0.6 (-0.6, 3.2) | |
| **baseline ctdna** | | | | **0.05** |
| Mean (sd) | 335.8 (673.9) | 197.6 (418.5) | 360.3 (709.4) | |
| Median (Min,Max) | 47 (0, 4475) | 1.9 (0.0, 1174.7) | 77.7 (0.1, 4475.0) | |
| **change ctdna group** | | | | **0.02** |
| Decrease from baseline | 33 (45) | 9 (82) | 24 (39) | |

| Predictors | Full Sample (n=73) | Censored (n=11) | Non Censored (n=62) | p-value |
|---|---|---|---|---|
| Increase from baseline | 40 (55) | 2 (18) | 38 (61) | |

*Table 1: Summary Statistics of the pembrolizumab data*

*(**l size** = target lesion size, **pdl1** = PD L1 %, **tmb** = TMB/mutations)*

Due to computational times and complexities, only the OSD algorithm, CQR and Robust Cox were run. Readers are encouraged to explore BHT, DBHT and other algorithms to see if they come to a consensus.

Coefficients and the concordance c-index for a Cox Proportional Hazards model with and without outlier detection using OSD are presented below.

Cox Proportional Hazards

| | coef | exp(coef) | se(coef) | z | Pr(>\|z\|) |
|---|---|---|---|---|---|
| age | -0.03 | 0.97 | 0.01 | -2.58 | 0.01 |
| sexMale | 0.40 | 1.50 | 0.37 | 1.10 | 0.27 |
| cohortB | 1.09 | 2.98 | 0.60 | 1.82 | 0.07 |
| cohortC | 0.87 | 2.38 | 0.61 | 1.43 | 0.15 |
| cohortD | -0.60 | 0.55 | 0.63 | -0.94 | 0.35 |
| cohortE | -0.45 | 0.64 | 0.44 | -1.01 | 0.31 |
| l_size | 0.00 | 1.00 | 0.00 | -0.63 | 0.53 |
| pdl1 | -0.01 | 0.99 | 0.01 | -1.61 | 0.11 |
| tmb | -0.16 | 0.85 | 0.14 | -1.15 | 0.25 |
| baseline_ctdna | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 |
| change_ctdna_groupIncrease from baseline | 0.93 | 2.54 | 0.31 | 2.99 | 0.00 |

Cox Proportional Hazards (removed outliers)

| | coef | exp(coef) | se(coef) | z | Pr(>\|z\|) |
|---|---|---|---|---|---|
| age | -0.01 | 0.99 | 0.01 | -0.73 | 0.46 |
| sexMale | 0.89 | 2.43 | 0.44 | 2.03 | 0.04 |
| cohortB | 1.78 | 5.93 | 0.75 | 2.38 | 0.02 |
| cohortC | 2.20 | 9.06 | 0.73 | 3.04 | 0.00 |
| cohortD | -0.24 | 0.79 | 0.76 | -0.31 | 0.75 |
| cohortE | -0.17 | 0.84 | 0.52 | -0.33 | 0.74 |
| l_size | -0.01 | 0.99 | 0.00 | -2.04 | 0.04 |
| pdl1 | -0.01 | 0.99 | 0.01 | -1.43 | 0.15 |
| tmb | -0.47 | 0.62 | 0.22 | -2.15 | 0.03 |
| baseline_ctdna | 0.00 | 1.00 | 0.00 | 1.80 | 0.07 |
| change_ctdna_groupIncrease from baseline | 1.71 | 5.51 | 0.39 | 4.36 | 0.00 |

| *Concordance C-index* | *Cox PH* | *Cox PH with OSD* |
|---|---|---|
| | 0.757 | 0.827 |

We see that, after running OSD, which removed 5 outlying observations, most predictors were statistically significant at the 0.05 level vs. the model without OSD. For example, Sex now has a p-value of 0.04 rather than 0.27.

More importantly, we saw an increase in the concordance c-index from 0.757 to 0.827, which is close to 1. These results are very promising as the removal of outlying observations allows the model to better capture the underlying relationship. As an alternative, estimates are also provided for a Robust Cox Regression based on Bednarski's method.

Robust Cox Regression

| variables | coefficients |
|---|---|
| age | -0.01 |
| sexMale | 0.38 |
| cohortB | 1.25 |
| cohortC | 1.10 |
| cohortD | -0.61 |
| cohortE | -0.48 |
| l_size | 0.00 |
| pdl1 | -0.01 |
| tmb | -0.28 |
| baseline_ctdna | 0.00 |
| change_ctdna_groupIncrease from baseline | 1.55 |

Below are the results of Censored Quantile Regression (CQR) for the 20th and 40th percentiles. As expected, the coefficients differ for different quantiles of the response variable. Although the concordance c-index is not available for CQR, this should not be an issue as the heterogeneity of observations were considered by the model.

CQR 20th percentile

| | Coefficient | CI Lower | CI Upper | SE | Test Statistic | P-value |
|---|---|---|---|---|---|---|
| (Intercept) | 3.12 | 0.91 | 5.80 | 1.25 | 2.50 | 0.01 |
| age | 0.00 | -0.05 | 0.02 | 0.02 | -0.07 | 0.95 |
| sexMale | 0.00 | -1.01 | 0.76 | 0.45 | 0.01 | 0.99 |
| cohortB | -1.32 | -2.81 | 1.93 | 1.21 | -1.09 | 0.27 |
| cohortC | -1.24 | -4.03 | 1.79 | 1.49 | -0.84 | 0.40 |
| cohortD | 0.63 | -1.50 | 4.70 | 1.58 | 0.40 | 0.69 |
| cohortE | 0.51 | -0.75 | 2.00 | 0.70 | 0.73 | 0.47 |
| l_size | 0.00 | 0.00 | 0.02 | 0.01 | 0.21 | 0.83 |
| pdl1 | 0.02 | 0.00 | 0.05 | 0.01 | 1.68 | 0.09 |
| tmb | 0.02 | -0.66 | 1.31 | 0.50 | 0.04 | 0.97 |
| baseline_ctdna | 0.00 | 0.00 | 0.00 | 0.00 | -0.03 | 0.97 |
| change_ctdna_groupIncrease from baseline | -1.84 | -3.39 | 0.36 | 0.96 | -1.92 | 0.05 |

CQR 40th percentile

| | Coefficient | CI Lower | CI Upper | SE | Test Statistic | P-value |
|---|---|---|---|---|---|---|
| (Intercept) | 3.63 | -2.07 | 9.56 | 2.97 | 1.22 | 0.22 |
| age | -0.02 | -0.10 | 0.04 | 0.04 | -0.47 | 0.64 |
| sexMale | 0.15 | -1.56 | 1.42 | 0.76 | 0.19 | 0.85 |
| cohortB | -0.83 | -3.73 | 2.77 | 1.66 | -0.50 | 0.62 |
| cohortC | -0.03 | -2.62 | 2.66 | 1.35 | -0.02 | 0.98 |
| cohortD | 11.21 | -18.82 | 24.52 | 11.05 | 1.01 | 0.31 |
| cohortE | 0.79 | -1.58 | 3.81 | 1.37 | 0.58 | 0.56 |
| l_size | 0.00 | -0.02 | 0.02 | 0.01 | 0.37 | 0.71 |
| pdl1 | 0.06 | -0.07 | 0.30 | 0.10 | 0.61 | 0.54 |
| tmb | 0.95 | -3.27 | 5.80 | 2.31 | 0.41 | 0.68 |
| baseline_ctdna | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.85 |
| change_ctdna_groupIncrease from baseline | -2.50 | -7.74 | 0.42 | 2.08 | -1.20 | 0.23 |

## 4. Conclusion and Discussion

Throughout this paper, various remedies were introduced to deal with the biases introduced by outlying observations in the Cox Proportional Hazards model. We found that dealing with outliers using OSD changed the estimates and significance of covariates.

Furthermore, the predictive performance of the model improved as the concordance c-index improved from 0.757 to 0.827. Applying OSD will allow future researchers to continue adopting the Cox model in their work. Censored Quantile Regression is another possible solution as the regression coefficients $\beta(\theta)$ differed for different quantiles ($\theta$) of the survival time. Care is needed as CQR is modelling the event time directly, rather than the hazard ratio.

From our results, OSD has been shown to be quite promising. One limitation is that we need prior knowledge of the maximal number of outliers we wish to remove. For example, if there are only 3 observations that are considered outliers, we certainly do not want to specify the removal of 5 outliers. This would cause two observations that are useful to the estimation to be deleted, which reintroduces biases into the model. Based on the literature, a general rule of thumb is to set the upper threshold at 10% of the dataset.

For CQR, there are many available estimation algorithms [5], which leads to varying results. Perhaps more importantly, the concordance c-index is not available for this method, which does not allow us to assess the predictive capabilities. Furthermore, one subtle point that was not introduced in our work are variable selection techniques, as different combinations of predictors will lead to varying regression estimates. To gain an improved sense of the outlier detection algorithms presented, further simulation studies are recommended.

Code:

The R code used can be found at https://github.com/entekang/CHL5250/blob/main/remedies%20of%20outlying%20observations%20in%20survival%20analysis.Rmd

References

[1] E. Carrasquinha, A. Verissimo, S. Vinga, Consensus outlier detection in survival analysis using the rank product test (2018).

[2] J. Pinto, A. Carvalho, S. Vinga, Outlier Detection in Survival Analysis based on the Concordance C-index (2015).

[3] J. Pinto, A. Carvalho, S. Vinga, Outlier detection in Cox proportional hazards models based on the concordance c-index (2015).

[4] X. Xue, X. Xie, H. Strickler, A censored quantile regression approach for the analysis of time to event data, Statistical Methods in Medical Research 27 (3) (2018) (955-965)

[5] A. Yazdani, et al., The comparison of censored quantile regression methods in prognosis factors of breast cancer survival