View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Essex Research Repository

Submitted

EFFICIENT EMPIRICAL LIKELIHOOD INFERENCE FOR RECOVERY RATE OF COVID-19 UNDER DOUBLE-CENSORING

BY JIEHU[1], WEILIANG[2,*], HONGSHENGDAI[3] AND YANCHUN BAO[4]

[1]*School of Mathematical Science, Xiamen University, China.* hujiechelsea@xmu.edu.cn

[2]*School of Mathematical Science, Xiamen University, China. Perelman School of Medicine, University of Pennsylvania, USA.*
[*]wei.liang@pennmedicine.upenn.edu; wliang@xmu.edu.cn

[3]*Department of Mathematical Sciences, University of Essex, UK.* hdaia@essex.ac.uk

[4]*Department of Mathematical Sciences, University of Essex, UK.* ybaoa@essex.ac.uk

Doubly censored data are very common in epidemiology studies. Ignoring censorship in the analysis may lead to biased parameter estimation. In this paper, we highlight that the publicly available COVID19 data may involve high percentage of double-censoring and point out the importance of dealing with such missing information in order to achieve better forecasting results. Existing statistical methods for doubly censored data may suffer from the convergence problems of the EM algorithms or may not be good enough for small sample sizes. This paper develops a new empirical likelihood method to analyse the recovery rate of COVID19 based on a doubly censored dataset. The efficient influence function of the parameter of interest is used to define the empirical likelihood (EL) ratio. We prove that $-2\log(\text{EL-ratio})$ asymptotically follows a standard χ^2 distribution. This new method does not require any scale parameter adjustment for the log-likelihood ratio and thus does not suffer from the convergence problems involved in traditional EM-type algorithms. Finite sample simulation results show that this method provides much less biased estimate than existing methods, when censoring percentage is large. The method application to the COVID19 data will help researchers in other field to achieve better estimates and forecasting results.

1. Introduction.

Doubly censored data, with both right and left censoring, occur when time-to-event data are censored either from above or below. Doubly-censored data are very common in studies of infectious disease with incubation period. The left censoring happens when the originating date of the incubation period is not fully observed due to practical sampling factors beyond experimental control. The date of the failure event is often right-censored. A particular doubly censored data on AIDS study can be found in De Gruttola and Lagakos (1989). Another example is time from symptom onset to recovery for people who get COVID19. For COVID19 studies (Verity et al., 2020), the incubation rate and recovery rate are the key factors for us to understand the epidemiology. In particular, in the current COVID19 outbreak, better understanding of the recovery rate will help governments to take the right intervention strategy at the right time. However, many existing research for COVID19 are based on published information from government or ministry of health websites and media reports (Verity et al., 2020). Such data have high percentage of missing information, such as high percentage of left or right censoring. This may distort the estimation of recovery rate, which could further distort the epidemiology model forecasting, as we can see from (Ferguson et al., 2020) that different model parameters will give very different forecasting results.

The dataset used in Verity et al. (2020) is from

Keywords and phrases: COVID19, Doubly censored data, Efficient influence function, Empirical likelihood

https://github.com/mrc-ide/COVID19_CFR_submission

which has a large number of missing information on the symptom onset and on the date of recovery. Our main research interests here are to study the recovery time, e.g. the time from symptom onset to recovery X, and to study the sensitivity of recovery rate on the epidemiology forecasting. The recovery times are clearly observed under right censoring because when the data were reported, recovery may not have happened to many patients. Therefore the right censoring time R is the time from the symptom onset date to the reporting date. On the other hand the left censoring time L is from the date of exposure to the date of recovery. As we know that the symptom can only occur after exposure to the virus, when symptom onset date is missing but the date of exposure to virus is available, we can retrieve information of X that $X \le L$.

In summary, under doubly censoring, the event time X of a subject cannot be observed unless it falls in an "observation interval" $[L, R]$. We observe L in the case of left censoring with $X < L$, or observe R in the case of right censoring with $X > R$. Let $(X_i,\, L_i,\, R_i), i = 1, \cdots, n$, be n independent copies of $(X,\, L,\, R)$, then observations under doubly censorship can be summarized as n independent pairs $(W_i,\, \delta_i), i = 1, \cdots, n$, where

$$W_i = \max(\min(X_i,\, R_i),\, L_i), \text{ and } \delta_i = \begin{cases} 1, & \text{if } L_i \le X_i \le R_i, \\ 2, & \text{if } X_i > R_i, \\ 3, & \text{if } X_i < L_i. \end{cases}$$

Usually, we assumed that the failure time X is independent of censoring vector $(L,\, R)$.

Denote F as the cumulative distribution function of X. Suppose that we are interested in a parameter θ, defined by a functional $\theta = \theta(F)$. Many important parameters can be represented as this form or, sometimes, we obtain θ via the corresponding estimating equation $g(X, \theta)$. For example, if we are interested in the expectation of a known function $m(X)$, then $\theta = \int m(x)\,\mathrm{d}F(x)$, and the corresponding estimating equation is $g(X,\, \theta) = m(X) - \theta$. Other examples include:

[1.] θ is the cumulative hazard function at given time t_0, i.e. $\theta = -\ln(1 - F(t_0))$, then the estimating equation is $g(X,\, \theta) = I_{\{X > t_0\}} - e^{-\theta}$;

[2.] θ is the mean residual life time at given time t_0, i.e.

$$\theta = E(X - t_0 | X > t_0) = \bar{F}^{-1}(t_0) \int_{t_0}^{\infty} (s - t_0)\,\mathrm{d}\,F(s),$$

where $\bar{F} = 1 - F$, then the estimating equation is $g(X,\, \theta) = (X - t_0 - \theta) I_{\{X \ge t_0\}}$.

To draw inference on the unknown parameter θ, a straightforward approach is to implement a distribution function estimation for F (Turnbull and Crowley, 1974; Tsai and Crowley, 1985; Chang and Yang, 1987; Chang, 1990)). Using the distribution function estimation, the asymptotic-normality based confidence interval for the parameter of interest θ can be constructed via the asymptotic variance estimator of the parameter estimate. But there are two main drawbacks associated with this method. First, the asymptotic variance usually takes a complicated form. Secondly, these confidence intervals based on asymptotic normal distribution do not always perform well for small samples. Other existing research about doubly-censored data may depend on specific model assumptions, such as (quantile) regression analysis (Zhang and Li, 1996; Ren and Gu, 1997; Ji et al., 2012) and two-sample tests (Shen et al., 2016). In this paper, we will solve these estimation problems via empirical likelihood method (Owen, 1988), which is a very useful tool for constructing confidence regions for θ in nonparametric settings. In general, the empirical likelihood approach has a number of advantages, such as the shape of the confidence region is determined automatically by the data. In many cases, the log empirical likelihood ratio statistics has asymptotic χ^2 distribution, therefore the confidence interval for θ can be constructed without estimating asymptotic variance.

Based on estimating equation $g(X, \theta)$, the original Empirical Likelihood (OEL) in Owen (1988) is defined as

$$\mathcal{R}^O(\theta) = \sup\left\{\prod_{i=1}^n np_i \Big| \sum_{i=1}^n p_i\, g(X_i, \theta) = 0,\ \sum_{i=1}^n p_i = 1,\ p_i \geq 0,\ i = 1, 2, \cdots, n\right\}.$$

It can be proved that

$$\mathcal{L}^O(\theta_0) = -2\log \mathcal{R}^O(\theta_0) \to \chi^2(1), \text{ in dist.}$$

Such empirical likelihood approach has many desirable statistical properties and some advantages over other competitors such as the normal approximations and the bootstrap methods. A very important work by Qin and Lawless (1994) generalized the EL method to make inference for parameter defined by a general estimating equation. Choosing different estimating equation to define likelihood statistics $\mathcal{R}^O(\theta)$ will lead to different inference process.

However, applying OEL methods to incomplete data will lead to a scaled χ^2 result. When the data is right censored, Wang and Jing (2001) utilized the Buckley-James estimator to define the estimating equation, and proved that the asymptotic distribution of the corresponding log-likelihood is a scaled χ^2 distribution. This limiting distribution can be used to construct the confidence interval for θ, if the scaled parameter is estimated. To avoid estimating the scaled parameter, He et al. (2016) used the efficient influence function of the parameter under right censorship to define the log-likelihood ratio statistics and proved its asymptotic distribution is a χ^2 distribution. The confidence interval for θ based on this method is much more accurate. Under doubly censoring, Ren (2001) proposed Leveraged Bootstrap Empirical Likelihood (LBEL) by combining the EL method with the bootstrap. Since the asymptotic distribution of the log-likelihood based on LBEL method is a scaled χ^2 distribution, the scaled parameter as an adjustment coefficient needs to be estimated in practice. Besides, the LBEL method demands that the parameter of interest should be the linear functional of F.

Notice that the EL likelihood function $\prod\limits_{i=1}^n p_i$ is not the real likelihood function for doubly censored data, Murphy and van der Vaart (1997) defined the likelihood function based on observations $\{(W_i, \delta_i)\}_{i=1}^n$

$$\mathcal{L}^{DC}(F) = \prod_{i=1}^n \Delta F(W_i)^{I_{\{\delta_i=1\}}} (\bar{F}(W_i))^{I_{\{\delta_i=2\}}} F(W_i)^{I_{\{\delta_i=3\}}}, \tag{1}$$

where DC is the abbreviation for Double Censoring, $\Delta F(t) = F(t) - F(t-)$ and $\bar{F} = 1 - F$ is the survival function. Using (1), they showed that this log-likelihood ratio subject to nonparametric moment constraints obeys the Wilks' phenomenon under some assumptions. This method avoids the scaled parameter, but is computationally difficult to find the nonparametric maximum likelihood. To solve this problem, Shen et al. (2016) proposed an EM algorithm to calculate this log-likelihood ratio statistics. However, EM algorithm may suffer from the problem of convergence to a local maximum point. Different from Shen et al. (2016), we investigate another approach in this paper. Inspired by He et al. (2016), we develop the likelihood statistics defined by efficient score function for the parameter of interest θ. This method is called Efficient-EL method in our paper. Under this new approach, we demonstrate that the log empirical likelihood ratio converges to the standard χ^2 distribution without using any scale parameter adjustment, which means the confidence interval for different kinds of parameters θ can be obtained by a unified algorithm. In the mean time, it is computationally much more efficient than existing EL methods under doubly censoring.

The rest of the paper is organized as follows. The Efficient-EL inference for the differential functional parameter θ under doubly-censored data is given in Section 2, including the large sample properties and the computing algorithm. Simulation studies of the Efficient-EL and the EM-EL method proposed by Shen et al. (2016) are provided in Section 3. We find that our approach performs much better for longer tail distributions, which usually lead to higher censoring proportions. In the mean time, the new method still performs as good as existing methods for lighter tail distributions which lead to lower censoring proportions. An application on COVID-19 study based on our proposed methodology is presented in Section 4. The paper concludes with a discussion in Section 5.

2. Efficient Empirical Likelihood Inference.

Denote $G_L(t) = \mathrm{P}\{L \le t\}$ and $G_R(t) = \mathrm{P}\{R \le t\}$ as the distribution of L and R respectively. Suppose we are interested in the estimation problem for a parameter $\theta = \theta(F)$, and the corresponding estimating equation for θ is $g(X, \theta)$, that means $\mathrm{E}\, g(X, \theta) = 0$. Since X cannot be observed unless it falls in $[L, R]$, we define

$$g^{DC}(W, \delta; \theta) = I_{\{\delta=1\}} \frac{g(W,\theta)}{G_R(W) - G_L(W)} + I_{\{\delta=2\}} \frac{g(W,\theta)}{1 - G_R(W)} + I_{\{\delta=3\}} \frac{g(W,\theta)}{G_L(W)}.$$

It is easy to see that, given the distribution F, G_L, G_R, we have $\mathrm{E}\, g^{DC}(W, \delta; \theta) = 0$ which gives an estimating equation for θ. Then, the EL ratio can be defined by

$$\mathcal{R}^{DC}(\theta) = \sup \left\{ \prod_{i=1}^{n} np_i \Big| \sum_{i=1}^{n} p_i\, g^{DC}(W_i, \delta_i; \theta) = 0, \sum_{i=1}^{n} p_i = 1,\, p_i \ge 0,\, i = 1, \cdots, n \right\}.$$

Substituting the unknown G_L, G_R with its consistent estimators will lead to a scaled asymptotic χ^2 distribution. Murphy and van der Vaart (1997) used the likelihood function (1) to solve the problem. Different from their idea, we will try to reconsider the estimating equation to overcome the scaled χ^2 asymptotic distribution problem.

2.1. *The main theorems.*

Assume $[\alpha, \beta] \subset [0, \infty)$ be the support of F, and the following assumptions hold.

(A1) $$G_L(x) - G_R(x-) > 0 \text{ on } x \in [\alpha, \beta],$$

(A2) $$F, G_L \text{ and } G_R \text{ are continuous with } G_L(\beta) = 1, G_R(\alpha) = 0.$$

Define $\mathrm{BV}[\alpha, \beta] = \{h : [\alpha, \beta] \to \mathbb{R},\, h \text{ is bounded and of bounded variation}\}$ and $H_F = \{h \in \mathrm{BV}[\alpha, \beta] : \int h\, \mathrm{d}F = 0\}$. The following Lemma provides the efficient influence function for θ.

LEMMA 2.1. *Let $dF_t(x) = (1 + t\, h(x))\, dF(x)$ be a submodel of $F(x)$, which approaches F at direction $h \in H_F$. Assume (A1) and (A2) hold and the Hadamard derivative of $\theta(F_t)$ exists, denoted by $\dot\theta_0$. Then the efficient influence function for θ is*

$$\psi(w, \delta; \theta) = \ell_F (\ell^* \ell_F)^{-1} \dot\theta_0,$$

where ℓ_F is the score operator

$$(\ell_F h)(w, \delta) = I_{\{\delta=1\}} h(w) + I_{\{\delta=2\}} \frac{\int_{(w,\infty)} h\, dF}{1 - F(w)} + I_{\{\delta=3\}} \frac{\int_{[0,w]} h\, dF}{F(w)},$$

and ℓ^* is its corresponding adjoint operator

$$(\ell^* g)(s) = g(s, 1)\Big(G_L(s) - G_R(s-)\Big) + \int_{[0,s)} g(u, 2)\, dG_R(u) + \int_{[s,\infty)} g(u, 3)\, dG_L(u).$$

PROOF. See Appendix . □

The assumptions (A1) and (A2) guarantee the operator $\ell^*\ell_F : \mathrm{BV}[\alpha, \beta] \to \mathrm{BV}[\alpha, \beta]$ is invertible. The following are some examples of derivatives $\dot{\theta}_0$ (in all of the examples we let t_0 be fixed).

[1.] For mean $\theta = \mathrm{E}\, X$, we have $\dot{\theta}_0 = x - \theta$.
[2.] For the kth moments $\theta = \mathrm{E}\, X^k$, we have $\dot{\theta}_0 = x^k - \theta$.
[3.] For cumulative distribution function $\theta = F(t_0)$, we have $\dot{\theta}_0 = \mathrm{I}_{\{x \le t_0\}} - \theta$.
[4.] For cumulative hazard function $\theta = -\ln(1 - F(t_0))$, we have

$$\dot{\theta}_0 = 1 - \mathrm{e}^{\theta}\, \mathrm{I}_{\{x > t_0\}}.$$

Since the operators ℓ^* and ℓ_F dependent on (F, G_L, G_R), we should write $\psi = \psi(w, \delta; \theta, F, G_L, G_R)$ more precisely. Let $\xi = (F, G_L, G_R)$, the efficient influence function can be denote as $\psi(W, \delta; \theta, \xi)$, hence

$$\mathrm{E}\, \psi(W, \delta; \theta, \xi) = 0.$$

Notice that the nuisance parameter ξ is unknown, we need to estimate it firstly.

For $j = 1, 2, 3$, define

$$\hat{H}_k(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\{W_i \le t,\, \delta_i = k\}} \quad \text{and} \quad \hat{H}(t) = \sum_{k=1}^{3} \hat{H}_k(t).$$

Chang and Yang (1987) gave the self-consistent estimators $\hat{F}$, $\hat{G}_L$, $\hat{G}_R$ of F, G_L, G_R by solving the following equations:

$$\hat{H}(t) = (1 - \hat{F}(t))\hat{G}_R(t) + \hat{F}(t)\hat{G}_L(t), \tag{2}$$

$$\hat{G}_R(t) = \int_0^t \frac{\mathrm{d}\hat{H}_2(u)}{1 - \hat{F}(u)}, \tag{3}$$

$$\hat{G}_L(t) = 1 - \int_t^\infty \frac{\mathrm{d}\hat{H}_3(u)}{\hat{F}(u)}. \tag{4}$$

Based on equation (2), a naive and simple iterative algorithm can be used to get $\hat{F}$, and then $\hat{G}_L$, $\hat{G}_R$ can be calculated by equations(3) and (4). In order to guarantee the asymptotic consistency and normality of $\hat{F}$, $\hat{G}_L$ and $\hat{G}_R$, we assume F, G_L and G_R satisfy conditions (A1)–(A6) in Chang (1990) throughout this paper.

Define $\hat{\xi} = (\hat{F}, \hat{G}_L, \hat{G}_R)$, then the efficient influence function $\psi(W_i, \delta_i; \theta, \xi)$ for θ can be estimated by $\psi(W_i, \delta_i; \theta, \hat{\xi})$. For simplicity of notations, denote $\psi_i(\theta) = \psi(W_i, \delta_i; \theta, \xi)$ and $\hat{\psi}_i(\theta) = \psi(W_i, \delta_i; \theta, \hat{\xi})$, then the corresponding Efficient EL ratio is defined as

$$\hat{\mathcal{R}}^{eDC}(\theta) \doteq \hat{\mathcal{R}}^{eDC}(\theta, \hat{\xi}) = \sup\left\{ \prod_{i=1}^{n} np_i \Big| \sum_{i=1}^{n} p_i\, \hat{\psi}_i(\theta) = 0,\, \sum_{i=1}^{n} p_i = 1,\, p_i \ge 0,\, i = 1, 2, \cdots, n \right\}. \tag{5}$$

Using Lagrangian multipliers, $p_i = 1/n(1 + \lambda\, \hat{\psi}_i(\theta))$, we further have

$$\hat{\mathcal{R}}^{eDC}(\theta) = \prod_{i=1}^{n} \frac{1}{1 + \lambda\, \hat{\psi}_i(\theta)},$$

where λ is the solution of the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\psi}_i(\theta)}{1+\lambda\,\hat{\psi}_i(\theta)}=0$$

and the following asymptotic results.

THEOREM 2.1. *Suppose the assumptions in Lemma 2.1 hold, θ_0 is the true value of the parameter of interest, and* $\boldsymbol{E}\,\psi^2(W,\,\delta;\,\theta_0)$ *exists, then we have*

$$\hat{\mathcal{L}}^{eDC}(\theta_0)\equiv -2\log\hat{\mathcal{R}}^{eDC}(\theta_0)\to\chi^2(1),\quad \textit{in dist.}$$

PROOF. Under the Lemma B.1 and Lemma B.2 in Appendix, this proof is similar to the proof of Original EL and therefore it is omitted. □

Theorem 2.1 shows that the estimated log empirical likelihood ratio converges to the standard χ^2 distribution without adjustment, which means the confidence interval for different kinds of parameters θ can be obtained by an unified algorithm. Hence, a confidence region for the parameter θ with asymptotic coverage probability $1-\alpha$ can be define as

$$CI=\left\{\theta:\,2\sum_{i=1}^{n}\log\left(1+\lambda\,\hat{\psi}_i(\theta)\right)\le\chi^2_\alpha(1)\right\}. \tag{6}$$

By recalling the definition of the efficient influence function for θ

$$\psi(w,\,\delta;\,\theta)=\ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0,$$

in the following subsection we present an algorithm for the calculation of the numerical solution of $\hat{\psi}_i$ and the confidence region CI.

2.2. *Algorithm for Efficient-EL Method.* Before presenting the algorithm, we need to introduce the following notations,

$$\hat{K}_1(t)=\begin{cases} n^{-1}\sum_{i=1}^{n}\left(1-\hat{F}(W_i)\right)^{-2}I_{\{\delta_i=2,W_i<t\}}, & \text{if}\quad t<B_n,\\ \hat{K}_1(B_n-), & \text{if}\quad t\ge B_n,\end{cases}$$

$$\hat{K}_2(t)=\begin{cases} n^{-1}\sum_{i=1}^{n}\hat{F}^{-2}(W_i)I_{\{\delta_i=3,W_i\ge t\}}, & \text{if}\quad t\ge A_n,\\ \hat{K}_2(A_n), & \text{if}\quad t<A_n,\end{cases}$$

where $A_n=\min\left\{W_i:\,\hat{F}(W_i)>0\right\}$, $B_n=\max\left\{W_i:\,\hat{F}(W_i-)<1\right\}$, and

$$K_{ij}=\frac{1}{n}\frac{\hat{K}_1(W_i\wedge W_j)+\hat{K}_2(W_i\vee W_j)}{\hat{G}_L(W_j)-\hat{G}_R(W_j-)}\,I_{\{\delta_j=1\}}.$$

For a given θ, define the least favorable direction $h_\theta(x)=(\ell^*\ell_F)^{-1}\dot{\theta}_0(x;\,\theta)$, then the efficient influence function is $\psi(w,\,\delta;\,\theta,\,\xi)=\ell_F h_\theta(x)$. Notice that only the values of $\psi(w,\,\delta;\,\theta,\,\xi)$ at the sample points $(W_i,\,\delta_i)$ are needed, therefore we can just calculate $h_\theta(W_1),\,h_\theta(W_2),\,\cdots,\,h_\theta(W_n)$. The following Corollary 2.1 shows a key equation for $\hat{h}_\theta(W_i)$ which will be used in the Efficient-EL algorithm.

COROLLARY 2.1. *The estimator $\hat{h}_\theta(W_i)$ satisfies the equation*

$$\begin{pmatrix} \dot{\theta}_0(W_1;\theta) \\ \vdots \\ \dot{\theta}_0(W_n;\theta) \end{pmatrix} = \begin{pmatrix} \Delta G_1 + K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & \Delta G_2 + K_{22} & \cdots & K_{2n} \\ \cdots & & & \\ K_{n1} & K_{n2} & \cdots & \Delta G_n + K_{nn} \end{pmatrix} \begin{pmatrix} \hat{h}_\theta(W_1) \\ \vdots \\ \hat{h}_\theta(W_n) \end{pmatrix}, \tag{7}$$

where $\Delta G_i := \hat{G}_L(W_i) - \hat{G}_R(W_i-)$.

The Efficient-EL ratio $\hat{\mathcal{R}}^{eDC}(\theta)$ can then be calculated by the following algorithm. Hence, the confidence interval for θ, CI in (6), can be constructed using the output $\hat{\psi}_i(\theta)$ by this algorithm.

Algorithm 1 Efficient-EL Algorithm

```
Solving (2)-(4) to get the self-consistent estimators $\hat{F}$, $\hat{G}_L$ and $\hat{G}_R$ of $F$, $G_L$ and $G_R$.
for $i = 1$ to $n$, do
    Calculate $\dot{\theta}_0(W_i;\theta)$ and $\Delta G_i = \hat{G}_L(W_i) - \hat{G}_R(W_i-)$,
    for $j = 1$ to $n$, do
        Calculate $K_{ij}$.
    end for
end for
Solve the equation (7) and get $\hat{h}_\theta(W_1)$, $\hat{h}_\theta(W_2)$, $\cdots$ , $\hat{h}_\theta(W_n)$.
for $i = 1$ to $n$, do
    Calculate $\hat{\psi}_i(\theta) = \psi(W_i, \delta_i; \theta, \hat{\xi})$,
    if $A_n \le W_i < B_n$ then
        $\hat{\psi}_i(\theta) = I_{\{\delta_i=1\}}\hat{h}_\theta(W_i) + \frac{I_{\{\delta_i=2\}}}{n(1-\hat{F}(W_i))}\sum_{k=1}^{n}\frac{I_{\{\delta_k=1,\,W_k>W_i\}}\hat{h}_\theta(W_k)}{\Delta G_k} + \frac{I_{\{\delta_i=3\}}}{n\hat{F}(W_i)}\sum_{k=1}^{n}\frac{I_{\{\delta_k=1,\,W_k\le W_i\}}\hat{h}_\theta(W_k)}{\Delta G_k}$,
    else if $W_i < A_n$ then
        $\hat{\psi}_i(\theta) = \psi(A_n, \delta_i; \theta, \hat{\xi})$,
    else
        $\hat{\psi}_i(\theta) = \psi(B_n-, \delta_i; \theta, \hat{\xi})$.
    end if
end for
Output $\hat{\psi}_i(\theta)$.
```

3. Simulation Studies.

In this section, we will illustrate the performance of our method via simulation studies. Here we compare our Efficient-EL method with EM-EL method in Shen et al. (2016).

In our simulations, we denote Exp(λ) as the exponential distribution with mean λ, Normal(μ, σ^2) as the normal distribution with mean μ and variance σ^2, Weibull(a,b) as the Weibull distribution with scale parameter a and shape parameter b, and Uniform(a,b) as the uniform distribution on $[a,b]$. For a given sample size n, we generate doubly-censored observations (W_i, δ_i). Based on the simulated data, we use all complete data X_i to construct the benchmark confidence interval, and compare it with the Efficient-EL confidence interval proposed in the previous section and EM-EL confidence interval (Shen et al., 2016).

TABLE 1
Coverage probabilities for Mean under Uniform$(0,\,3)$ distribution

		Nominal Level = 0.90			Nominal Level = 0.95		
		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete	0.898	0.894	0.896	0.944	0.943	0.945
	Efficient	0.897	0.888	0.877	0.944	0.935	0.932
	EM	0.896	0.878	0.873	0.944	0.932	0.930
n=80	Complete	0.906	0.899	0.897	0.955	0.949	0.948
	Efficient	0.903	0.892	0.887	0.950	0.942	0.940
	EM	0.904	0.891	0.882	0.951	0.943	0.940
n=100	Complete	0.909	0.896	0.903	0.953	0.948	0.954
	Efficient	0.906	0.898	0.899	0.951	0.946	0.944
	EM	0.905	0.888	0.891	0.952	0.944	0.942

Two percentages in each column stand for left censoring proportion and right censoring proportion.

TABLE 2
Coverage probabilities for MRL(t_0) under Uniform$(0,\,3)$ distribution

		Nominal Level = 0.90			Nominal Level = 0.95		
t_0 = 0.1 quantile of F		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete	0.907	0.898	0.906	0.954	0.943	0.953
	Efficient	0.904	0.874	0.854	0.949	0.929	0.910
	EM	0.898	0.851	0.831	0.947	0.914	0.894
n=80	Complete	0.895	0.894	0.892	0.949	0.948	0.945
	Efficient	0.892	0.880	0.866	0.941	0.931	0.921
	EM	0.886	0.859	0.835	0.936	0.921	0.907
n=100	Complete	0.896	0.889	0.897	0.949	0.948	0.947
	Efficient	0.896	0.884	0.868	0.949	0.936	0.923
	EM	0.898	0.859	0.838	0.946	0.925	0.907
t_0 = 0.5 quantile of F		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete	0.897	0.891	0.896	0.945	0.943	0.950
	Efficient	0.871	0.838	0.819	0.928	0.896	0.877
	EM	0.888	0.833	0.831	0.938	0.897	0.895
n=80	Complete	0.895	0.901	0.891	0.949	0.950	0.841
	Efficient	0.885	0.856	0.844	0.935	0.912	0.909
	EM	0.889	0.848	0.846	0.941	0.909	0.915
n=100	Complete	0.893	0.901	0.897	0.949	0.948	0.947
	Efficient	0.892	0.873	0.871	0.942	0.928	0.923
	EM	0.894	0.857	0.864	0.945	0.921	0.929

Two percentages in each column stand for left censoring proportion and right censoring proportion.

3.1. *Simulation Results for Mean and Mean Residual Lifetime.* In this simulation, there are two parameters we are interested. One is the mean of X, denoted by θ_1, therefore the estimating equation for θ_1 is $g_1(X,\theta_1) = X - \theta_1$. The other is the Mean Residual Lifetime (MRL) of X given t_0, denoted by $\theta_2(t_0)$, the corresponding estimating equation is $g_2(X,\theta_2) = (X - t_0 - \theta_2)I_{\{X \geq t_0\}}$.

Uniform$(0,3)$ and Normal$(0,1)$ distributions are considered as the underlying lifetime distribution F. When X follows the uniform distribution, the left censoring time L and censoring interval length $R-L$ are uniformly distributed on interval $[c_1,c_2]$ and $[c_3,c_4]$. When X is drawn from the normal distribution, the left censoring time L and censoring interval length $R-L$ are distributed as Normal$(\mu_1,\,1)$ and Normal $(\mu_2,\,1)$. We set c_i and μ_i to be different values to achieve 10%, 20%, 30% left censoring proportions and 10%, 20%, 30% right censoring proportions respectively. Based on 5000 simulated data sets, we construct Efficient-EL confidence intervals, EM-EL confidence intervals and Complete EL confidence intervals. The coverage probabilities are summarized in Table 1 - Table 3.

From the results in these Tables, we noticed that as the sample size n increases, all coverage probabilities converge to the nominal levels. When n is fixed, coverage probabilities of Efficient-EL confidence intervals and EM-EL confidence intervals decrease as the censoring

TABLE 3
Coverage probabilities for Mean and MRL(t_0) under Normal$(0, 1)$ distribution

		Nominal Level = 0.90			Nominal Level = 0.95		
		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete	0.897	0.887	0.898	0.945	0.938	0.943
	Efficient	0.880	0.854	0.850	0.932	0.918	0.913
	EM	0.887	0.873	0.885	0.939	0.935	0.938
n=80	Complete	0.899	0.898	0.900	0.950	0.948	0.948
	Efficient	0.884	0.870	0.867	0.939	0.929	0.928
	EM	0.894	0.881	0.896	0.942	0.937	0.948
n=100	Complete	0.911	0.905	0.899	0.957	0.951	0.948
	Efficient	0.895	0.877	0.870	0.949	0.931	0.928
	EM	0.903	0.882	0.893	0.955	0.937	0.946
t_0 = 0.1 quantile of F		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete	0.899	0.890	0.890	0.948	0.942	0.943
	Efficient	0.871	0.850	0.836	0.929	0.914	0.905
	EMEL	0.885	0.871	0.846	0.941	0.928	0.909
n=80	Complete	0.899	0.895	0.893	0.949	0.946	0.946
	Efficient	0.888	0.871	0.863	0.942	0.931	0.925
	EMEL	0.896	0.882	0.866	0.946	0.929	0.923
n=100	Complete	0.896	0.902	0.905	0.951	0.954	0.954
	Efficient	0.888	0.875	0.867	0.938	0.934	0.927
	EMEL	0.895	0.889	0.862	0.945	0.941	0.924

Two percentages in each column stand for left censoring proportion and right censoring proportion.

proportion increases. The coverage probabilities of the confidence intervals for parameter MRL(t_0) decrease when t_0 increases. In all cases, the performance of Efficient-EL and EM-EL methods are close to that of Complete EL method when censoring proportion is not large.

Under uniform distribution, Efficient-EL and EM-EL methods perform similarly in the coverage probability estimation for the parameter mean (see Table 1). The difference among these two methods and Complete EL method is small, especially for small censoring proportion or large sample size. However, the performance of these methods for the parameter MRL is different (see Table 2). The coverage probabilities of Efficient-EL confidence intervals performs better than that of EM-EL almost for all scenarios when $t_0 = 0.1$. Meanwhile, Efficient-EL method performs as good as EM-EL when $t_0 = 0.5$ for most cases. Under the scenario with normal distribution (see Table 3), both EM-EL method and Efficient-EL method provide very good coverage probabilities.

3.2. *The impact of different censoring proportions – exponential distribution.* In this section, we investigate the impact of different censoring proportions. Here we still consider the parameter mean θ_1 and $\theta_2(t_0)$ =MRL(t_0), where t_0 is the 30% quantile of the distribution F, Exp(1). We set the left censoring time L as Exp(c_1) and censoring interval length $R - L$ as Exp(c_2). We choose c_i to be different values to achieve left 20% right 40%, left 30% right 30% and left 40% right 20% censoring proportions respectively. Based on 5000 simulated data sets, the coverage probabilities are summarized in Table 4.

When right censoring proportion is large (40%), from Table 4, we can see that the coverage probabilities of Efficient-EL are much better than that of EM-EL, although both methods have larger bias. Besides, the coverage probabilities of Efficient-EL confidence intervals converge faster to the nominal level than that of EM-EL. Take the parameter MRL as an example, as the sample size increase from 50 to 500, the coverage probabilities of Efficient-EL increase from 0.637 to 0.753, while EM-EL only increase from 0.666 to 0.710.

3.3. *The impact of different censoring proportions – lognormal distribution.* We still consider the same parameters of interests, mean θ_1 and MRL $\theta_2(t_0)$, t_0 = 30% quantile of distribution F, which is LogNorm$(0, 0.64)$ in this subsection. The left censoring time L

TABLE 4
Coverage probabilities for Mean and MRL under Exp(1) distribution

		Nominal Level = 0.90			Nominal Level = 0.95		
Mean		20%+40%	30%+30%	40%+20%	20%+40%	30%+30%	40%+20%
n=50	Complete	0.881	0.884	0.879	0.932	0.938	0.932
	Efficient	0.692	0.762	0.819	0.762	0.835	0.883
	EM	0.688	0.813	0.880	0.768	0.879	0.939
n=100	Complete	0.892	0.893	0.898	0.949	0.939	0.946
	Efficient	0.736	0.800	0.846	0.807	0.865	0.910
	EM	0.712	0.823	0.898	0.795	0.888	0.946
n=150	Complete	0.892	0.889	0.894	0.944	0.945	0.943
	Efficient	0.748	0.799	0.836	0.818	0.865	0.906
	EM	0.728	0.827	0.886	0.809	0.893	0.941
n=500	Complete	0.896	0.897	0.899	0.944	0.946	0.948
	Efficient	0.786	0.835	0.862	0.852	0.899	0.924
	EM	0.731	0.840	0.884	0.808	0.908	0.936
MRL		20%+40%	30%+30%	40%+20%	20%+40%	30%+30%	40%+20%
n=50	Complete	0.869	0.873	0.876	0.922	0.928	0.929
	Efficient	0.637	0.724	0.798	0.710	0.799	0.869
	EM	0.666	0.787	0.862	0.739	0.858	0.921
n=100	Complete	0.885	0.879	0.895	0.938	0.939	0.947
	Efficient	0.692	0.770	0.826	0.763	0.846	0.887
	EM	0.688	0.812	0.870	0.764	0.879	0.929
n=150	Complete	0.894	0.895	0.901	0.948	0.939	0.949
	Efficient	0.720	0.781	0.827	0.797	0.855	0.899
	EM	0.707	0.817	0.879	0.784	0.879	0.935
n=500	Complete	0.899	0.897	0.903	0.948	0.947	0.947
	Efficient	0.753	0.827	0.859	0.830	0.894	0.920
	EM	0.710	0.827	0.884	0.792	0.893	0.938

Two percentages in each column stand for left censoring proportion and right censoring proportion.

follows Exp(c_1), and censoring interval length $R - L$ follows LogNorm$(c_2, 0.25)$. Let c_i to be different values to achieve 20% left censoring and 40% right censoring, 30% left censoring and 30% right censoring, and 40% left censoring and 20% right censoring, respectively. Based on 5000 simulated data sets, the coverage probabilities are summarized in Table 5.

Based on Table 5, similar to the results for exponential distribution example, we can see that higher right censoring proportion leads to lower coverage probabilities. The coverage probabilities of confidence intervals constructed by the proposed Efficient-EL approach is much better than EM-EL methods for large sample sizes. Although coverage probabilities for both methods have a bias because of the large censoring percentages, we can still see that the coverage probabilities of Efficient-EL based confidence intervals steadily increase, while the coverage probabilities of EM-EL seem not to have a clear increasing pattern (coverage probabilities of EM-EL may not converge to the nominal level as sample size becomes larger). This means that the new Efficient-EL approach is more reliable for heavy-tailed distributions and for highly-censored data.

4. Analysis of COVID19 Data.

There has already been a vast literature on COVID19 research. A widely-used mathematical model is the Susceptible-Exposed-Infectious-Resistant (SEIR) epidemiology model, based on which the UK government's lock-down strategy were made Ferguson et al. (2020). It models the flows of people between the four states: susceptible (S), exposed (E), infected (I), and resistant (R) via

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta}{N}SI, \quad \frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\beta}{N}SI - \sigma E,$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I, \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I,$$

TABLE 5
Coverage probabilities for Mean and MRL under LogNorm$(0,\, 0.64)$ distribution

		Nominal Level = 0.90			Nominal Level = 0.95		
Mean		20%+40%	30%+30%	40%+20%	20%+40%	30%+30%	40%+20%
n=50	Complete	0.853	0.872	0.873	0.915	0.923	0.925
	Efficient	0.585	0.737	0.802	0.662	0.808	0.865
	EM	0.624	0.821	0.855	0.706	0.894	0.917
n=100	Complete	0.881	0.875	0.878	0.938	0.929	0.932
	Efficient	0.599	0.742	0.823	0.680	0.824	0.883
	EM	0.628	0.826	0.842	0.711	0.893	0.901
n=150	Complete	0.877	0.893	0.876	0.936	0.944	0.932
	Efficient	0.611	0.768	0.835	0.699	0.838	0.886
	EM	0.623	0.824	0.830	0.712	0.892	0.890
n=500	Complete	0.895	0.896	0.890	0.948	0.948	0.943
	Efficient	0.605	0.779	0.841	0.693	0.854	0.905
	EM	0.552	0.788	0.793	0.638	0.861	0.871
MRL		20%+40%	30%+30%	40%+20%	20%+40%	30%+30%	40%+20%
n=50	Complete	0.843	0.849	0.841	0.907	0.910	0.907
	Efficient	0.525	0.688	0.784	0.595	0.770	0.849
	EM	0.566	0.789	0.845	0.655	0.862	0.904
n=100	Complete	0.873	0.876	0.877	0.925	0.930	0.929
	Efficient	0.542	0.731	0.818	0.624	0.804	0.880
	EM	0.563	0.808	0.834	0.647	0.878	0.900
n=150	Complete	0.877	0.879	0.878	0.933	0.934	0.933
	Efficient	0.555	0.740	0.805	0.628	0.813	0.870
	EM	0.574	0.802	0.820	0.634	0.873	0.888
n=500	Complete	0.895	0.899	0.892	0.941	0.947	0.945
	Efficient	0.574	0.769	0.822	0.657	0.836	0.891
	EM	0.529	0.773	0.779	0.614	0.850	0.853

Two percentages in each column stand for left censoring proportion and right censoring proportion.

In this SEIR model, the infectious rate β controls the rate of spread which represents the probability of transmitting disease between a susceptible and an infectious individual. The incubation rate σ is the rate of latent individuals becoming infectious (average duration of incubation is $1/\sigma$). Recovery rate γ is determined by the average duration of infection. $N = S + E + I + R$ is the total population. As in Ferguson et al. (2020) we assume people recovered from the disease is immune to it. The basic reproductive number, $R_0 = \beta/\gamma$, does not change in this model.

Recovery rate γ is a very important parameter in such SEIR model. When estimating this parameter, publicly available data could have a large proportion of missing information. For example, the dataset from

https://github.com/mrc-ide/COVID19_CFR_submission

has a large number of missing information on the symptom onset and on the date of recovery. It actually gave a double censoring dataset for the recovery time. The event time X of interest is time length from symptom onset to recovery. The right censoring variable R is from symptom onset to the reporting date. The left censoring variable L is from the date of exposure to recovery. The total number of observations used in our analysis is $n = 547$ and the data are collected from 20th January 2020 to 28th February 2020.

Firstly, we list the censoring proportions of this dataset under different groups in Table 6. Using the Efficient-EL method we proposed, the confidence intervals of recovery time for different groups can be calculated. These results also list in Table 6. From Table 6, we can see that the elder groups have longer average recovery period, but there is no significant different between male and female.

We also carry out a simulation study similar to Ferguson et al. (2020) to compare the forecasting results based on different model parameter values, in order to address the importance

TABLE 6

The analysis of COVID19 data for different groups

Group	left	observed	right	sample size	CI Lower	Mean	CI Upper
Male	0.052	0.185	0.763	323	17.370	19.842	22.153
Female	0.063	0.184	0.753	218	18.171	20.243	21.567
Age under 30	0.140	0.215	0.645	85	9.527	17.759	22.411
Age 30-50	0.082	0.212	0.707	186	17.651	19.605	21.172
Age 50-60	0.066	0.168	0.766	115	18.947	21.731	23.813
Age 60-70	0.076	0.124	0.800	83	17.902	22.041	24.627
Age over 70	0.089	0.089	0.822	68	18.951	22.173	25.614
Overall	0.059	0.170	0.771	547	18.784	20.013	20.928

of parameter estimation for such forecasting analysis. In our simulation we set population $N = 10^6$ and discrete time steps of size $b = 0.1$ of a simulated day. At each step, the number of new exposed is drawn from a Poisson distribution with rate $\beta SIb/N$, the number of individuals becoming infectious or recovered are Poisson random variable with rates σEb and γIb, respectively. According to Novel Coronavirus (2019-nCoV) Situation Report-7, the incubation period is betwen 2 and 10 days. We therefore set $\sigma = 1/5.1$, the same as Ferguson et al. (2020). Since SEIR model dose not include mortality, we classify death and recovered as one group, re-estimate the recovery time and get the 95% confidence interval $[18.784, 20.928]$ and mean 20.013. Hence, three different recovery periods: short duration 15 days ($\gamma = 1/15$, corresponding to results without using double censoring analysis, no right censoring, over estimation of recovery rate), medium duration 20 days ($\gamma = 1/20$, corresponding to our result based on double censoring) and long duration 25 days ($\gamma = 1/25$, corresponding to results without using double censoring analysis, no left censoring, under estimation of recovery rate) are considered in our simulation.

We also consider two different quarantine protocols: no government interventions $R_0 = 2.4$ following Ferguson et al. (2020) and with mild government interventions $R_0 = 1.5$, which lead to the parameter value $\beta = R_0\gamma$ in our simulation. All of our simulation are carried out via the R package deSolve of SEIR model. The daily new cases are plotted in Figure 1, where for the curves from left to right, the dashed line means 15-day recovery period, the solid line means 20-day recovery period and the dotted line means 25-day recovery period. For both $R_0 = 2.4$ and $R_0 = 1.5$ we can see that with a shorter recovery time, the COVID19 outbreak will end much quicker. Also the mode of daily infected cases will be much smaller under the scenario of shorter recovery time.

To achieve the herd immunity proposed by the UK government requires a proportion of the UK population being immune to the virus to stop it from spreading. It is well-known that such herd immunity can be stimulated by vaccination or recovery following infection. Based our result using a sophisticated double censoring statistical model, we can see clearly that the recovery period should be much shorter than the estimated figures proposed by other existing works. With $R_0 = 1.5$, the peak of the curve with recovery rate $1/20$ is will occur on day 592 (95% confidence interval $[527, 761]$), the peak with recovery rate $1/15$ will occur on day 479 (95% confidence interval $[422, 609]$) and the peak with recover rate $1/25$ will occur on day 705 (95% confidence interval $[621, 883]$). Therefore, with a slight over or under estimation for the recovery rate, the forecasting peak date will be different at a scale of about 110 days. This would imply that the outbreak could end about four months earlier than people expected.

5. Conclusions. Through our COVID19 forecasting analysis and Ferguson et al. (2020), we can see that correct estimation of SEIR model parameters may change the final forecasting results significantly, for example the peak date estimation may be different at the scale of months. For such a rapid spread disease, it will be extremely challenging to carry out real-time monitor the pandemic (Birrell et al., 2020). The data collected in real-time will certainly

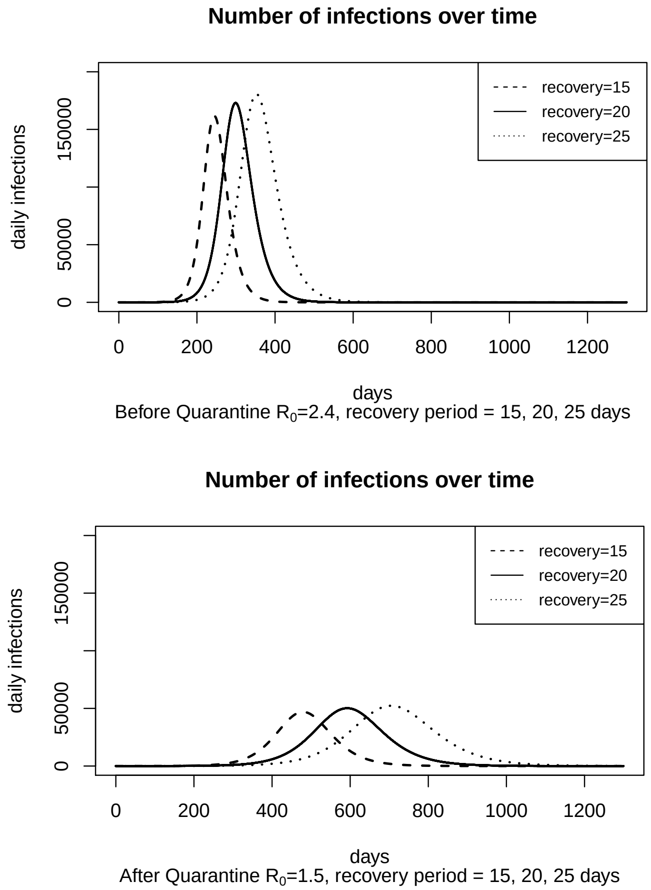

FIG 1. *Increased infections curves before and after quarantine. Three different sets of curves represent different recovery period from left to right.*

involve different kind of censoring. This paper highlighted the importance of dealing with the censored data and presented a efficient new statistical estimation approach. By utilizing the efficient influence function of the parameter of interest as an estimating equation, a new method of constructing EL confidence interval for doubly censored data is proposed in this paper. This new Efficient-EL method is easy to calculate since it does not need to estimate scale parameter. Simulation studies show that the new method performances better than the EM-EL method in terms of coverage probabilities.

Comparing model predictions with our estimated recovery rate parameter and existing parameter values used in other research works, we found that the peak of the epidemic predicted could be months different from each other. This could lead to wrong health policy decisions, for example taking or removing lock-down decisions at the wrong time points, which may lead to a second peak of outbreak or making the lock-down period too long to cause severe economic damage and mental health problems for more people. Our analysis highlights the importance of doing such sophisticated survival analysis will provide better estimation for the parameters in the SEIR models.

To our knowledge, this is the first work which considered using censoring techniques in survival analysis to carry out parameter estimation for COVID19 data. Most existing COVID19 research such as Kucharski et al. (2020) and Ferguson et al. (2020) did not address the issues of highly contaminated data due to censoring or simply use prespecified model parameters. Although only a relatively small data set is used, the methodology can be used by other researcher who have the access to larger COVID19 dataset with individual information. It will help interdisciplinary collaboration between statisticians and epidemiologists and help policy makers on public health policy making.

APPENDIX A: PROOF OF LEMMA 2.1

PROOF. For any $h \in H_F$, define $\mathrm{d}F_t = (1+th)\,\mathrm{d}F$, then the likelihood of doubly censored random variable is

$$L_t(w,\,\delta) = \Delta F_t(w)^{I_{\{\delta=1\}}}\,(\bar{F}_t(w))^{I_{\{\delta=2\}}}\,F_t(w)^{I_{\{\delta=3\}}}.$$

Let P_F be the distribution of doubly censored random variable $(W,\,\delta)$, then the score operator $\ell_F : H_F \to \mathbb{L}^2(P_F)$ is

$$\begin{aligned}(\ell_F h)(w,\,\delta) &= \frac{\partial}{\partial t}\Big|_{t=0} \ln L_t(w,\,\delta)\\ &= I_{\{\delta=1\}}h(w) + I_{\{\delta=2\}}\frac{\int_{(w,\,\infty)} h\,\mathrm{d}F}{1-F(w)} + I_{\{\delta=3\}}\frac{\int_{[0,\,w]} h\,\mathrm{d}F}{F(w)}.\end{aligned}$$

By the definition of the adjoint operator $\ell^* : \mathbb{L}^2(P_F) \to H_F$, for any $h_1,\,h_2 \in H_F$,

$$< \ell_F h_1,\,\ell_F h_2 >_{P_F} = < h_1,\,\ell^*\ell_F h_2 >_F .$$

Using Fubini's theorem,

$$\begin{aligned}&< \ell_F h_1,\,\ell_F h_2 >_{P_F} = \int (\ell_F h_1\,\ell_F h_2)\,\mathrm{d}P_F\\ &= \int h_1(x)h_2(x)\Big(G_L(x) - G_R(x-)\Big)\mathrm{d}F + \int \frac{\int_{(r,\infty)} h_1\mathrm{d}F \int_{(r,\infty)} h_2\mathrm{d}F}{1-F(r)}\mathrm{d}G_R(r)\\ &\qquad + \int \frac{\int_{[0,\,l]} h_1\mathrm{d}F \int_{[0,\,l]} h_2\mathrm{d}F}{F(l)}\mathrm{d}G_L(r)\\ &= \int h_1(x)\left(h_2(x)\Big(G_L(x) - G_R(x-)\Big) + \int_{[0,x)} \frac{\int_{(r,\infty)} h_2\mathrm{d}F}{1-F(r)}\mathrm{d}G_R(r) + \int_{[x,\,\infty)} \frac{\int_{[0,\,l]} h_2\mathrm{d}F}{F(l)}\mathrm{d}G_L(l)\right)\end{aligned}$$

we get

$$(\ell^*\ell_F h)(x) = \Big(G_L(x) - G_R(x-)\Big)h(x) + \int \left(\int_{[x\vee s,\,\infty)} \frac{\mathrm{d}\,G_L}{F} + \int_{[0,\,x\wedge s)} \frac{\mathrm{d}\,G_R}{1-F}\right) h(s)\,\mathrm{d}\,F(s).$$

According to Lemma A.2 (i) in Murphy and van der Vaart (1997), under the assumptions, the operator $\ell^*\ell_F$ is one to one, onto and continuously invertible. By the definition of $\dot{\theta}_0$, for any $h \in H_F$,

$$\frac{\partial}{\partial t}\Big|_{t=0}\theta(F_t) = \int \dot{\theta}_0\,h\,\mathrm{d}F = < \dot{\theta}_0,\,h >_F = < \ell^*\ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0,\,h >_F = < \ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0,\,\ell_F h >_{P_F} .$$

According to the definition in Tsiatis (2007), $\psi(w,\,\delta;\,\theta) = \ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0$ is the efficient influence function.

□

APPENDIX B: LEMMA FOR THEOREM 2.1

To prove the theorem 2.1, we need the following two Lemmas. Define

$$h_0 = (\ell^* \ell_F)^{-1} \dot{\theta}_0(x;\,\theta_0), \qquad \hat{h}_0 = (\ell^* \ell_{\hat{F}})^{-1} \dot{\theta}_0(x;\,\theta_0).$$

LEMMA B.1. *Under the assumptions of Theorem 2.1, we have*

$$\left|\left|\hat{h}_0 - h_0\right|\right|_\infty \to 0.$$

PROOF. From

$$\left|\left|\hat{h}_0 - h_0\right|\right|_\infty \le \left|\left|(\ell^* \ell_{\hat{F}})^{-1}((\ell^* \ell_F)h_0 - (\ell^* \ell_{\hat{F}})h_0)\right|\right|_\infty$$

we just need to prove

$$\left|\left|\ell^* \ell_F - \ell_* \ell_{\hat{F}}\right|\right|_\infty = \sup_{h \in \mathrm{BV}[\alpha,\,\beta]} \left|\left|\ell^* \ell_F h(x) - \ell^* \ell_{\hat{F}} h(x)\right|\right|_\infty \to 0, \tag{8}$$

and to prove the result $\left|\left|(\ell^* \ell_{\hat{F}})^{-1}\right|\right|_\infty$ is bounded.

[1.] Since

$$\ell_F h(w,\,\delta) = I_{\{\delta=1\}} h(w) + I_{\{\delta=2\}} \frac{\int_{(w,\,\infty)} h\,\mathrm{d}F}{1 - F(w)} + I_{\{\delta=3\}} \frac{\int_{[0,\,w]} h\,\mathrm{d}F}{F(w)},$$

so

$$\ell_F h(w,\,2) = \frac{\int_{(w,\infty)} h\,\mathrm{d}F}{1 - F(w)}, \qquad \ell_F h(w,\,3) = \frac{\int_{[0,\,w]} h\,\mathrm{d}F}{F(w)}.$$

Hence

$$\ell^* \ell_F h(x) = h(x)\Big(G_L(x) - G_R(x-)\Big) + \int_{[0,x)} \ell_F h(r,\,2)\,\mathrm{d}G_R(r) + \int_{[x,\,\infty)} \ell_F h(l,\,3)\,\mathrm{d}G_L(l),$$

and

$$\begin{aligned}
&\left|\left|\ell^* \ell_F h(x) - \ell^* \ell_{\hat{F}} h(x)\right|\right|_\infty = \Delta_1 + \Delta_2 + \Delta_3,\\
&\Delta_1 = \sup_{x \in [\alpha,\,\beta]} \left| \Big((G_L(x) - G_R(x-)) - (\hat{G}_L(x) - \hat{G}_R(x-)) \Big)\, h(x) \right|,\\
&\Delta_2 = \sup_{x \in [\alpha,\,\beta]} \left| \int_{[0,x)} \ell_F h(r,\,2)\,\mathrm{d}G_R(r) - \int_{[0,x)} \ell_{\hat{F}} h(r,\,2)\,\mathrm{d}\hat{G}_R(r) \right|,\\
&\Delta_3 = \sup_{x \in [\alpha,\,\beta]} \left| \int_{[x,\,\infty)} \ell_F h(l,\,3)\,\mathrm{d}G_L(l) - \int_{[x,\,\infty)} \ell_{\hat{F}} h(l,\,3)\,\mathrm{d}\hat{G}_L(l) \right|.
\end{aligned}$$

From Chang and Yang (1987), we have

$$\Delta_1 \le \sup_{x \in [\alpha,\,\beta]} \left| \Big(\hat{G}_L - G_L\Big)(x) h(x) \right| + \sup_{x \in [\alpha,\,\beta]} \left| \Big(\hat{G}_R - G_R\Big)(x-) h(x) \right| \to 0.$$

We also have

$$\begin{aligned}
\Delta_2 &= \sup_{x \in [\alpha,\beta]} \left| \int_{[0,x)} (\ell_{\hat{F}} h)(u,2)\,\mathrm{d}\hat{G}_R(u) - \int_{[0,x)} (\ell_F h)(u,2)\,\mathrm{d}G_R(u) \right| \\
&\le \sup_x \left| \int_{[0,x)} (\ell_{\hat{F}} h)(u,2)\,\mathrm{d}\left(\hat{G}_R(u) - G_R(u)\right) \right| + \sup_x \int_{[0,x)} \left| (\ell_{\hat{F}} h)(u,2) - (\ell_F h)(u,2) \right| \,\mathrm{d}G_R(u) \\
&\le \sup_x \left| \int_{[0,x)} (\ell_{\hat{F}} h)(u,2)\,\mathrm{d}\left(\hat{G}_R(u) - G_R(u)\right) \right| + \int \left| (\ell_{\hat{F}} h)(u,2) - (\ell_F h)(u,2) \right| \,\mathrm{d}G_R(u),
\end{aligned}$$

Lemma 3.1 in Chang (1990) shows that the first part of above equation is $o(1)$, and the proof of Lemma A.2(ii) in Murphy and van der Vaart (1997) shows that the second part is also $o(1)$. Therefore $\Delta_2 \to 0$. Similarly, we get $\Delta_3 \to 0$. Hence, for any $h \in \mathrm{BV}[\alpha,\beta]$,

$$\left|\left|\ell^*\ell_F h(x) - \ell^*\ell_{\hat{F}} h(x)\right|\right|_\infty \to 0.$$

[2.] Now we will prove that $\left|\left|(\ell^*\ell_{\hat{F}})^{-1}\right|\right|_\infty$ is bounded. For the convenience of proof, we denote $\mathcal{S} = \ell^*\ell_F$ and $\hat{\mathcal{S}} = \ell^*\ell_{\hat{F}}$. Next, we define

$$\hat{\mathcal{T}} = \mathcal{S}^{-1}(\mathcal{S} - \hat{\mathcal{S}}), \qquad \hat{\mathcal{U}} = \sum_{k=0}^{\infty} \hat{\mathcal{T}}^k.$$

It is easy to verify that

$$\hat{\mathcal{S}}^{-1} = \hat{\mathcal{U}}\,\mathcal{S}^{-1} \qquad \text{and} \qquad \hat{\mathcal{U}}^{-1} = I - \hat{\mathcal{T}}.$$

Since $||\hat{\mathcal{T}}||_\infty = ||\mathcal{S}^{-1}(\mathcal{S} - \hat{\mathcal{S}})||_\infty \le ||\mathcal{S}^{-1}||_\infty\, ||\mathcal{S} - \hat{\mathcal{S}}||_\infty \to 0$, we have

$$||\hat{\mathcal{U}}\,||_\infty = \left|\left|\sum_{k=0}^{\infty} \hat{\mathcal{T}}^k\right|\right|_\infty \le \sum_{k=0}^{\infty} ||\hat{\mathcal{T}}||_\infty^k = \frac{1}{1 - ||\hat{\mathcal{T}}||_\infty} \to 1.$$

Therefore

$$||\hat{\mathcal{S}}^{-1}||_\infty = ||\hat{\mathcal{U}}\,\mathcal{S}^{-1}||_\infty \le ||\hat{\mathcal{U}}\,||_\infty ||\mathcal{S}^{-1}||_\infty$$

is bounded.

□

LEMMA B.2. *Define* $\psi_{i0} = \psi_i(\theta_0)$, $\hat{\psi}_{i0} = \hat{\psi}_i(\theta_0)$ *and* $\sigma^2 = \boldsymbol{E}\,\psi^2(W,\,\delta;\,\theta_0)$. *Under the assumptions of Theorem 2.1, we have*

(1)
$$\max_{1\le i\le n} \left|\hat{\psi}_{i0}\right| = o_p(n^{1/2}),$$

(2)
$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_{i0}\right) \to N(0,\,\sigma^2),$$

(3)
$$\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_{i0}^2 = \sigma^2 + o_p(1).$$

PROOF. [1.] Since

$$\begin{aligned}\max_{1\le i\le n} \left|\hat{\psi}_{i0}\right| &\le \max_{1\le i\le n} \left|\hat{\psi}_{i0} - \psi_{i0}\right| + \max_{1\le i\le n} |\psi_{i0}| = \left(\max_{1\le i\le n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2\right)^{1/2} + \max_{1\le i\le n} |\psi_{i0}| \\ &\le n^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2\right)^{1/2} + o_p(n^{1/2}),\end{aligned}$$

we only need to prove $n^{-1}\sum_{i=1}^{n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2 = o_p(1)$. Note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2 &= \sum_{k=1}^{3}\left(\frac{1}{n}\sum_{i=1}^{n} \left|\psi(W_i,\,k;\,\theta_0,\,\hat{\xi}) - \psi(W_i,\,k;\,\theta_0,\,\xi)\right|^2 I_{\{\delta_i = k\}}\right) \\ &= \sum_{k=1}^{3} \int \left(\psi(w,\,k;\,\theta_0,\,\hat{\xi}) - \psi(w,\,k;\,\theta_0,\,\xi)\right)^2 \mathrm{d}\hat{H}_k(w) = \sum_{k=1}^{3} \Gamma_k,\end{aligned}$$

where

$$\Gamma_1 = \int \left(\psi(w,\,1;\,\theta_0,\,\hat{\xi}) - \psi(w,\,1;\,\theta_0,\,\xi)\right)^2 \mathrm{d}\hat{H}_1(w),$$

$$\Gamma_2 = \int \left(\psi(w,\,2;\,\theta_0,\,\hat{\xi}) - \psi(w,\,2;\,\theta_0,\,\xi)\right)^2 \left(1-\hat{F}(w)\right) \mathrm{d}\hat{G}_R(w),$$

$$\Gamma_3 = \int \left(\psi(w,\,3;\,\theta_0,\,\hat{\xi}) - \psi(w,\,3;\,\theta_0,\,\xi)\right)^2 \hat{F}(w)\,\mathrm{d}\hat{G}_L(w).$$

Using Lemma B.1, we have

$$\begin{aligned}\Gamma_1 &= \int \left(\ell_{\hat{F}}\hat{h}_0(w,\,1) - \ell_F h_0(w,\,1)\right)^2 \mathrm{d}\hat{H}_1(w) = \int \left(\hat{h}_0(w) - h_0(w)\right)^2 \mathrm{d}\hat{H}_1(w)\\ &\le \left|\left|\hat{h}_0 - h_0\right|\right|_\infty^2 \to 0.\end{aligned}$$

For the second part,

$$\begin{aligned}\Gamma_2 = &\int \left(\ell_{\hat{F}}\hat{h}_0(w,\,2) - \ell_F h_0(w,\,2)\right)^2 \left(1-\hat{F}(w)\right) \mathrm{d}\left(\hat{G}_R(w) - G_R(w)\right)\\ &+ \int \left(\ell_{\hat{F}}\hat{h}_0(w,\,2) - \ell_F h_0(w,\,2)\right)^2 \left(1-\hat{F}(w)\right) \mathrm{d}G_R(w).\end{aligned}$$

From Lemma 3.1 in Chang (1990), we know that the first part of above equation is $o(1)$. Since $\ell_{\hat{F}}\hat{h}_0(w,\,\delta) - \ell_F h_0(w,\,\delta) \to 0$, together with dominated convergence theorem, the second part of above equation is $o(1)$. Therefore $\Gamma_2 \to 0$, and Γ_3 converges to 0 can be proved similarly. Hence part (1) is proved.

[2.] Using the equations (2) - (4), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{\psi}_{i0} &= \sum_{k=1}^3 \int \psi(w,\,k;\,\theta_0,\,\hat{\xi})\,\mathrm{d}\hat{H}_k(w)\\ &= \int \hat{h}_0\left(\hat{G}_R - \hat{G}_L\right)\mathrm{d}\hat{F} + \int \left(1-\hat{G}_R\right)\hat{h}_0\,\mathrm{d}\hat{F} + \int \hat{G}_L\hat{h}_0\,\mathrm{d}\hat{F}\\ &= \int \hat{h}_0(w)\,\mathrm{d}\hat{F}(w) = \int \dot{\theta}_0\,\mathrm{d}\hat{F} = \int \dot{\theta}_0\,\mathrm{d}\left(\hat{F} - F\right).\end{aligned}$$

Due to the definition of efficient influence function and the proof of Lemma A.3 in Murphy and van der Vaart (1997), we have

$$\int \dot{\theta}_0\,\mathrm{d}(\hat{F} - F) = \frac{1}{n}\sum_{i=1}^n \ell_F h_0(W_i,\,\delta_i) + o_p(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \psi_{i0} + o_p(n^{-1/2}).$$

Therefore

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \hat{\psi}_{i0}\right) \to N(0,\,\sigma^2).$$

[3.] Since

$$\frac{1}{n}\sum_{i=1}^n \hat{\psi}_{i0}^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{\psi}_{i0} - \psi_{i0}\right)^2 + \frac{2}{n}\sum_{i=1}^n \left(\hat{\psi}_{i0} - \psi_{i0}\right)\psi_{i0} + \frac{1}{n}\sum_{i=1}^n \psi_{i0}^2,$$

and

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_{i0}-\psi_{i0}\right)\psi_{i0}\right| \le \left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_{i0}-\psi_{i0}\right)^2\right|^{1/2}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{i0}^2\right|^{1/2} = o_p(1),$$

we get

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_{i0}^2 = \frac{1}{n}\sum_{i=1}^{n}\psi_{i0}^2 + o_p(1) = \sigma^2 + o_p(1).$$

□

APPENDIX C: PROOF OF COROLLARY 2.1

PROOF. From the definition, the least favorable direction $h_\theta = (\ell^*\ell_F)^{-1}\dot{\theta}_0(x;\,\theta)$ satisfies the equation $(\ell^*\ell_F)h_\theta = \dot{\theta}_0(x;\,\theta)$, that is

(9)
$$\Big(G_L(x)-G_R(x-)\Big)h_\theta(x)+\int\left(\int_{[x\vee s,\,\infty)}\frac{\mathrm{d}\,G_L}{F}+\int_{[0,\,x\wedge s)}\frac{\mathrm{d}\,G_R}{1-F}\right)h_\theta(s)\,\mathrm{d}\,F(s)=\dot{\theta}_0(x;\,\theta).$$

Since

$$\mathrm{d}\,F(s)=\frac{\mathrm{d}\,H_1(s)}{G_L(s)-G_R(s-)},\quad \mathrm{d}\,G_R(s)=\frac{\mathrm{d}\,H_2(s)}{1-F(s)},\quad \mathrm{d}\,G_L(s)=\frac{\mathrm{d}\,H_3(s)}{F(s)}.$$

therefore

$$\begin{aligned}
&\int\left(\int_{[x\vee s,\,\infty)}\frac{\mathrm{d}\,G_L}{F}+\int_{[0,\,x\wedge s)}\frac{\mathrm{d}\,G_R}{1-F}\right)h_\theta(s)\,\mathrm{d}\,F(s)\\
&=\int\frac{1}{G_L(s)-G_R(s-)}\left(\int_{[0,\,x\wedge s)}\frac{\mathrm{d}\,H_2(u)}{(1-F(u))^2}+\int_{[x\vee s,\,\infty)}\frac{\mathrm{d}\,H_3(u)}{F^2(u)}\right)h_\theta(s)\,\mathrm{d}\,H_1(s)\\
&=\int\frac{1}{G_L(s)-G_R(s-)}\Big(K_1(x\wedge s)+K_2(x\vee s)\Big)h_\theta(s)\,\mathrm{d}\,H_1(s),
\end{aligned}$$

where

$$K_1(t)=\int_{[0,\,t)}\frac{\mathrm{d}\,H_2(u)}{(1-F(u))^2},\quad K_2(t)=\int_{[t,\,\infty)}\frac{\mathrm{d}\,H_3(u)}{F^2(u)}.$$

Hence, equation (9) can be rewritten as

(10)
$$\Big(G_L(x)-G_R(x-)\Big)h_\theta(x)+\int\frac{\Big(K_1(x\wedge s)+K_2(x\vee s)\Big)}{G_L(s)-G_R(s-)}h_\theta(s)\,\mathrm{d}\,H_1(s)=\dot{\theta}_0(x;\,\theta).$$

Substitute $\hat{F}$, $\hat{G}_R$ and $\hat{G}_L$ into K_1, K_2 and (10), we get $\hat{K}_1(t)$, $\hat{K}_2(t)$ and

(11)
$$\begin{aligned}
\dot{\theta}_0(x;\,\theta)&=\Big(\hat{G}_L(x)-\hat{G}_R(x-)\Big)h_\theta(x)\\
&\quad+\frac{1}{n}\sum_{j=1}^{n}\frac{\hat{K}_1(x\wedge W_j)+\hat{K}_2(x\vee W_j)}{\hat{G}_L(W_j)-\hat{G}_R(W_j-)}\,h_\theta(W_j)\,I_{\{\delta_j=1\}}.
\end{aligned}$$

Set $x=W_i$ $(i=1,\,2,\,\cdots,\,n)$ in equation (11), we get the equation (7).

□

REFERENCES

BIRRELL P. J., WERNISCH L., TOM B. D. M., HELD L., ROBERTS G. O., PEBODY R. G., and DE ANGELIS D. (2020). Efficient real-time monitoring of an emerging influenza pandemic: How feasible? *Annals of Applied Statistics.* **14**(1):74–93.

CHANG, M. N. (1990). Weak Convergence of Self-consistent Estimator of the Survival Function with Doubly Censored Data. *Annals of Statistics.* **18**, 391–404.

CHANG, M. N., YANG G. L.(1987). Strong Consistency of a Nonparametric Estimator of the Survival Function with Doubly Censored Data. *Annals of Statistics.* **15**, 1536–1547.

DE GRUTTOLA, V. and LAGAKOS, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1–12.

FERGUSON N. M., LAYDON D., NEDJATI-GILANI G., IMAI N., AINSLIE K., BAGUELIN M., BHATIA S., BOONYASIRI A., CUCUNUBA Z., CUOMO-DANNENBURG G., DIGHE A., DORIGATTI I., FU H., GAYTHORPE K., GREEN W., HAMLET A., HINSLEY W., OKELL L. C., VAN ELSLAND S., THOMPSON H., VERITY R., VOLZ E., WANG H., WANG Y., WALKER P. G. T., WALTERS C., WINSKILL P., WHITTAKER C., DONNELLY C. A., RILEY S. and GHANI A. C. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. The 9th report from the WHO Collaborating Centre for Infectious Disease Modelling. https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19

HE, S.Y., LIANG, W., SHEN, J.S. & YANG, G. (2016). Empirical Likelihood for Right Censored Lifetime Data. *Journal of Amreican Statistical Association.* **111**, 646–655.

JI, S., PENG, L., CHENG, Y. and LAI, H. (2012). Quantile Regression for Doubly Censored Data. *Biometrics.* **68**, 101–112.

KUCHARSKI, A. J., RUSSELL, T. W., DIAMOND, C., LIU, Y., EDMUNDS, J., FUNK, S. and EGGO, R. M. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet, Infectious Diseases*, **20**(5), May 2020, Pages 553–558.

MURPHY S. A., VAN DER VAART A. W.(1997). Semiparametric Likelihood Ratio Inference. *Annals of Statistics.* **25**, 1471–1509.

OWEN, A. B. (1988). Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika.* **75**, 237–249.

QIN, J., LAWLESS, L. (1994). Empirical Likelihood and General Estimating Equations. *Annals of Statistics.* **22**, 300–325.

REN, J.J. (2001). Weighted Empirical Likelihood Ratio Confidence Intervals for the Mean with Censored Data. *Annals of the Institute of Statistical Mathematics.* **53**, 498–516.

REN, J.J., GU, M.G. (1997). Regression M-estimators with doubly censored data. *The Annals of Statistics.* **25**, 2638–2664.

SHEN, J.S., YUEN K.C., LIU C.L. (2016). Empirical likelihood confidence regions for one- or two-samples with doubly censored data. *Computational Statistics and Data Analysis.* **93**, 285-293.

TSIATIS A.(2007). *Semiparametric theory and missing data.* Springer.

TSAI, W. Y., CROWLEY, J. (1985). A Large Sample Study of Generalized Maximum Likelihood Estimators from Incomplete Data via Self-Consistent. *Annals of Statistics.* **13**, 1317–1334.

TURNBULL, B. W., CROWLEY, J. (1974). Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association.* **69**, 169–173.

VERITY R., OKELL L. C., DORIGATTI I., WINSKILL P., WHITTAKER C., IMAI N., CUOMO-DANNENBURG G., THOMPSON H., WALKER P. G. T., FU H., DIGHE A., GRIFFIN J. T., BAGUELIN M., BHATIA S., BOONYASIRI A., CORI A., CUCUNUBA Z., FITZJOHN R., GAYTHORPE K., GREEN W., HAMLET A., HINSLEY W., LAYDON D., NEDJATI-GILANI G., RILEY S., VAN ELSLAND S., VOLZ E., WANG H., WANG Y. and XI X., (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet, Infectious Diseases*, March 30, 2020, DOI:https://doi.org/10.1016/S1473-3099(20)30243-7.

WANG, Q.H., JING, B.Y. (2001). Empirical Likelihood for a Class of Functions of Survival Distribution With Censored Data. *Annals of the Institute of Statistical Mathematics.* **53**, 517–527.

Novel Coronavirus (2019-nCoV) Situation Report-7 - World Health Organization (WHO), January 27, 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200127-sitrep-7-2019–ncov.pdf

ZHANG, C.H., LI, X. (1996). Linear regression with doubly censored data. *Annals of Statistics.* **24**, 2720–2743.