SELF-INTRODUCTION

Name :

Atina Husnaqilati

Educational Background:

- 2012-2016
  S.Si in Statistik, Universitas Gadjah Mada, Indonesia.
- 2017-2019
  M.Sc in Mathematics, Tohoku University, Japan.
- 2019-now
  Ph.D student in Mathematics, Tohoku University, Japan.

Publication:

- Husnaqilati, Atina. (2016). *Combining parametric, semi-parametric, and nonparametrik survival model with stacked method* http://etd.repository.ugm.ac.id/home/detail_pencarian/98064
- Salmahaminati, Salmahaminati & Husnaqilati, Atina & Yahya, Amri. (2017). Statistical t Analysis for the Solution of Prediction Trash Management in Dusun Tanjung Sari Kec. Ngaglik Kab Sleman, Yogyakarta. *Journal of Physics: Conference Series*. 795. 012046. 10.1088/1742-6596/795/1/012046.
- Husnaqilati, Atina & Utami, Herni & Danardono. (2018). Survival Analysis for Cancer Patient with Stacked Method. *Advanced Science Letters*. 24. 678-681. 10.1166/asl.2018.11786.

Internships:

- 1 month in National Family Planning Coordinating Agency Indonesia.

A predictive survival time for COVID-19 by stacked method

Husnaqilati, A.[1] Akama, Y.[1]

[1] Department of Mathematics, Tohoku University

数学・数理科学専攻若手研究者のための異分野・異業種研究交流会2021 13 November

Background

In late October 2021, the cases of COVID-19 began to decrease. However, we still need the precise mathematical models that capable of predicting the time to death which provide health officials with valuable information to develop appropriate strategies to reduce the death toll. Early studies have shown that statistical analysis to build predictive models can evaluate mortality rates which applied to COVID-19 issues.

Keywords: **COVID-19; survival function; Brier score; IPCW-Brier score, ROC.**

Method

To evaluate the predictive survival time of 1021 patients (Xu et al. 2020) by age, sex, and acute symptoms, we applied the stacked survival method (Wey et al. 2015) for building a new predictive model of survival time that combines different survival models:

1. log-normal model.
2. Cox PH.
3. Random survival forest (RSF).

To measure model performance, we employ the time-dependent area under the curve (AUC) receiver operating characteristic (ROC) (Heagerty, et al. 2002).

Aim

The focus of this study is to improve the prediction model of survival times from COVID-19 patients by age, sex, and acute symptoms.

Notation

For the sample size *n* and the covariates number *p*,

$\boldsymbol{x}_i$: *p*- dimensional covariate vector

δ_i : 0 (censored sample), 1 (uncensored sample)

T : time to death

C : time to censored

$S(t|\boldsymbol{x})$: the survival time function $P(T>t|\boldsymbol{x})$

$G(t|\boldsymbol{x})$: the survival function of cencored time $P(C>t|\boldsymbol{x})$

References

Wey, A., Connett, J., & Rudser, K. (2015). Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, **16**(3), 537-549.

Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., et al. (2020). Epidemiological data from the COVID-19 outbreak, real-time case information.

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56**(2)**, 337-344.

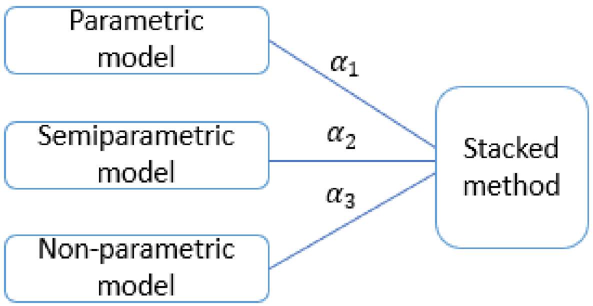
Lostritto, K., Strawderman, R. L., & Molinaro, A. M. (2012). A partitioning deletion/substitution/addition algorithm for creating survival risk groups. *Biometrics*, 68**(4)**, 1146-1156.

Contact email: Husqila@gmail.com

Stacked method

Stacked method

A stacked method for survival analysis combines multiple models of survival functions (Breiman 1996). The objectives of this method is to estimate *S(t|**x**)* from *m* candidate model.



The estimation of the survival functions obtains from the stacked model by *m* candidate models.

$$\hat{S}(t|\boldsymbol{x}) = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{m}} \hat{\alpha}_{\boldsymbol{k}}\hat{S}_{\boldsymbol{k}}(t|\boldsymbol{x})$$

where $\hat{\alpha}$ is the estimate of weights of all survival model α. To find $\hat{\alpha}$, we employ weighted least squares with constrain $\sum_{k=1}^{m}\alpha_k = 1$ and $\alpha_k \geq 0$.

Brier score

To solve $\hat{\alpha}$, we need the loss function of survival function. We apply Brier score to loss function of survival function as follows:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\{Z_i(t) - \hat{S}(t|\boldsymbol{x_i})\}^2.$$

Here, $Z_i(t)$ is indicator function $I(t_i > t)$ for $1 \leq i \leq n$.

IPCW-Brier score

Lostritto et al. (2012) improved Brier score for survival function with right censored. They introduced inverse probability of censoring weights (IPCW) as follows:

$$IPCW - BS(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t)|\boldsymbol{x}_i)}\{Z_i(t) - \hat{S}(t|\boldsymbol{x}_i)\}^2.$$

Here,

$T_i \quad = \min(\mathrm{C}, t_i, \mathrm{t})$

$\Delta_i(t) \quad = \delta_i\ (y_i \leq t); 1\ (y_i > t).$

Wey et al. (2015) estimated the survival function $G(.|\boldsymbol{x}_i)$ by Kaplan-Meier estimation.

Finally, by IPCW-Brier score, Wey et al. (2015) minimized the loss function over a set $t_1, t_2, \dots, t_s$. The estimation of weighted least square of α with $\sum_{k=1}^{m}\alpha_k = 1$ and $\alpha_k \geq 0$ for all *k*.

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\substack{\sum_{k=1}^{m}\alpha_k=1 \\ \alpha_k \geq 0\ (k=1,..,m)}} \sum_{r=1}^{s}\sum_{i=1}^{n}\frac{\Delta_i(t_r)}{\hat{G}(T_i(t_r)|\boldsymbol{x}_i)}\left\{Z_i(t_r) - \sum_{k=1}^{m}\alpha_k\,\hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)\right\}^2$$

Wey et al (2015) calculated $\hat{G}(T_i(t_r)|\boldsymbol{x}_i)$ by Kaplan-Meier estimation. Here $\hat{S}_k^{(-i)}(t_r|\boldsymbol{x}_i)$ is the *k*-th model's survival prediction for *n* samples from *i*-th observation during fitting process.

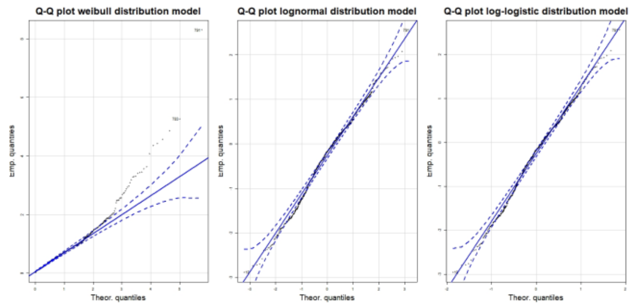
Survival function of COVID-19 data by stacked method

COVID-19 data analysis

1. The dataset contains cases of COVID-19 that recorded **from January 6th, 2020, to June 4th, 2020.**
2. The total number of patients is **1021** from 22 countries.
3. The outcome variable was **a survival time patient, constructed as the time between the date of confirmation and dead time**.
4. The censored samples are **patients with outcome discharge from hospital** because we do not know actual dead time.
5. The covariates are **age, sex, and acute symptoms**
6. We categorize the acute symptom which is **1 for a patient that has at least one of acute symptoms (acute pneumonia, acute cardiac or kidney injury, and acute respiratory distress syndrome (ARDS))** and 0 for others symptoms.

Fitting survival function

Before we apply the stacked method for the dataset, we assess the fitting of the COVID-19 dataset to some parametric probability survival distributions (Weibull distribution, lognormal distribution, log-logistic distribution), by using **Q-Q plot**.



we see that the dataset well fit **lognormal distribution or log-logistic**, because mostly the points plotted on the **graph lognormal and log-logistic lie on straight lines**.

Result of Stacked method

In the stacked survival models for COVID-19 dataset, we combine **log-normal model for a parametric model**, **Cox proportional hazard model (CoxPH) for a semi-parametric model**, and **random survival forests (RSFs) for a non-parametric model**.

	Variables	Coefficient	p-value	L95%	U95%	Alpha($\hat{\alpha}$)
Log-normal Model	Age	-0.0215	<2e-16	-0.0252	-0.0179	0.1516
	Sex	0.2066	0.0022	0.0745	0.3387	
	Acute symptoms	-0.5624	2.8e-06	-0.7977	-0.3271	
CoxPH	Age	0.0269	<2e-16	0.0227	0.0311	0.4208
	Sex	-0.2195	0.0028	-0.3638	-0.0752	
	Acute symptoms	0.5650	5.79e-06	0.3206	0.8092	
RSFs						0.4276

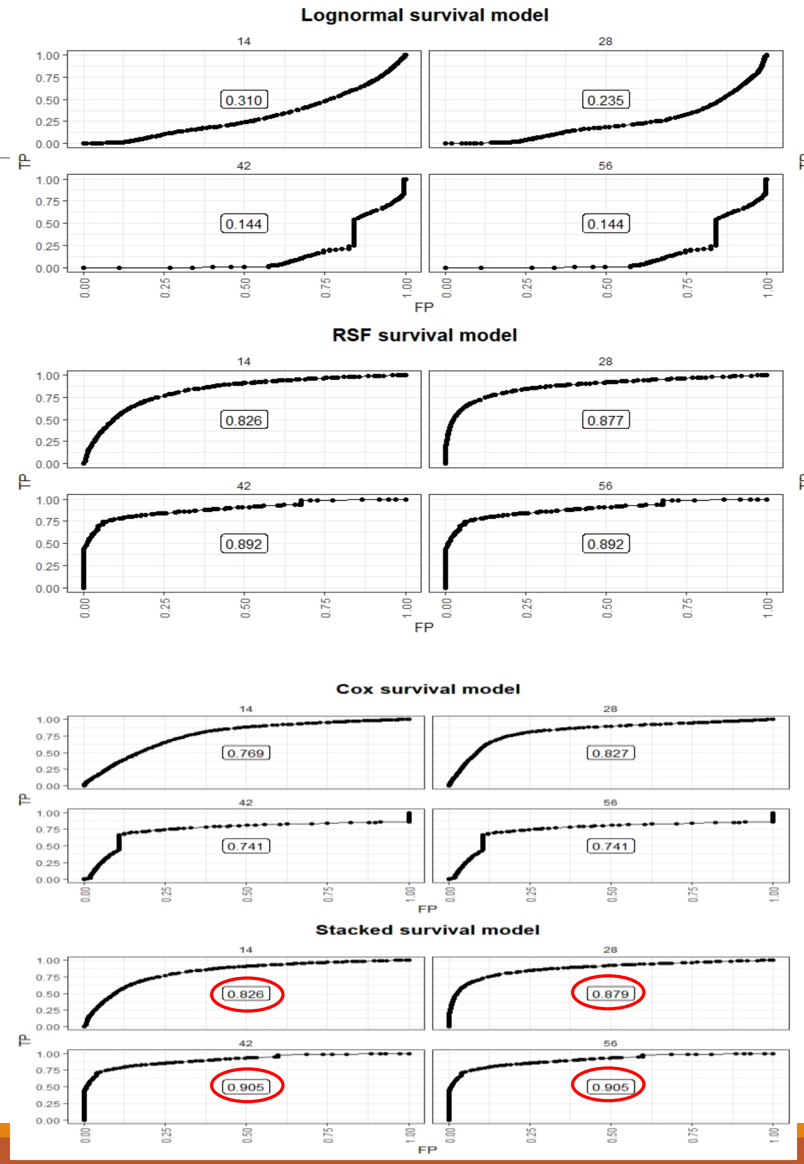
The above table shows the estimation of weighted least square $\hat{\alpha}$ **stacked model**. Moreover, we see that all variables (age, sex, acute symptoms) are significant covariates for three model survivals.

Stacked model performance and conclusion

Time-dependent area under the curve receiver operating characteristic (ROC)

To measure model performance, the stacked model was compared with the three survival models (Log-normal distribution, Cox Proportional hazard, RSFs) based on time-dependent area under the curve receiver operating characteristic (ROC)

The focus of the time-dependent area under the curve AUCs was on the 2 weeks to 4 weeks post confirmation of COVID-19 patients. By the figures in this slide, the stacked method is the largest AUC (the area under the ROC curve) for all selected specific *t* time. Therefore, **stacked method for survival model outperforms and has flexibility for time prediction.**



Conclusion

The stacked model improve the prediction model of survival times from COVID-19 patients by age group, sex at the different level, and acute symptoms based on time-dependent area under the curve receiver operating characteristic (ROC). This result provides a basis for health officials to develop appropriate strategies to reduce the death toll.