Mixture Model for Interval Censored Outcomes, in the presence of inflation of zeroes

Yibai Zhao†, Ryan Sheehan†, Raji Balasubramanian†

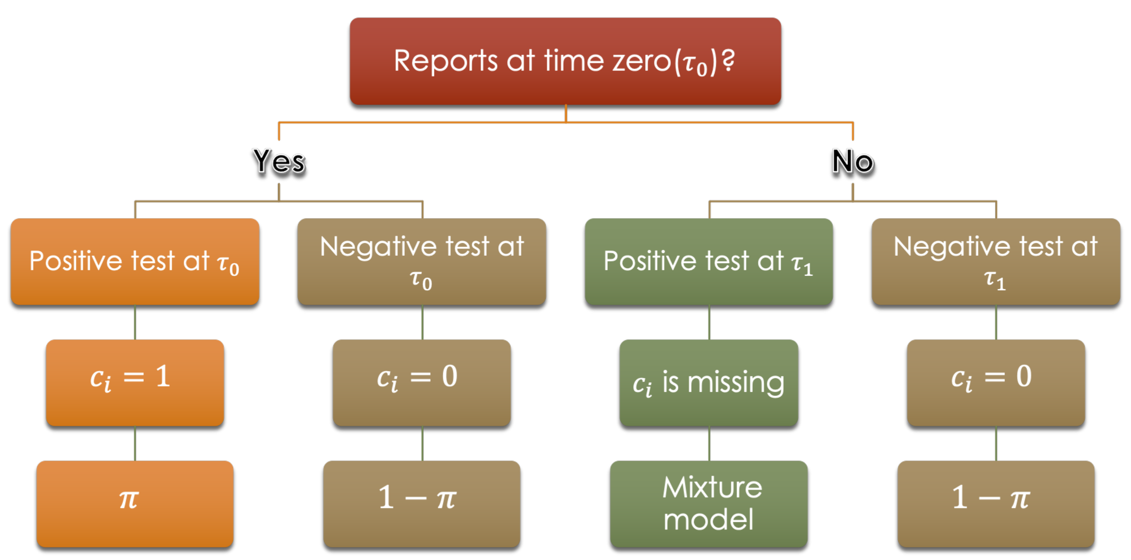
†Department of Biostatistics & Epidemiology, University of Massachusetts Amherst

UMassAmherst
School of Public Health & Health Sciences
Biostatistics and Epidemiology

Background and Motivation

- Silent events such as incident Type 2 diabetes cannot be observed directly. A sequentially diagnostic tests will be scheduled to ascertain the events.
- In the observational cohort settings, usually, there's a non-ignorable proportion of the participants already test positive for the outcome of interest at the beginning of the study.
- In [2], a stratified Weibull model was implemented to interval-censored outcome by relaxing the proportional hazard (PH) assumption across levels of a subset of explanatory data.
- In [1], a mixture model of logistic regression and Weibull survival model was proposed to account for the zero inflation at baseline in the interval-censored settings.
- We proposed a mixture model to evaluate the relationship between predictors and the interval-censored time to event of interest as well as accommodating the inflation of zeros at the baseline and allowing baseline hazard function to vary across strata defined by a subset of explanatory variables.

Tree Diagram of Mixture Model



Model

- Mixture model

  The true distribution of time to event T can be seen as an mixture distribution having mixture proportion $\pi(\boldsymbol{z};\boldsymbol{\theta_1})$ for positive tests at baseline and $(1-\pi(\boldsymbol{z};\boldsymbol{\theta_1}))$ for the rest of classes, with component distribution $Pr(T \le t;\boldsymbol{z}|\boldsymbol{c}=1)=1$ and $Pr(T \le t;\boldsymbol{z}|\boldsymbol{c}=0)=1-S(t;\boldsymbol{\theta}_2|\boldsymbol{c}=0,\boldsymbol{z})$. So the final mixture model can be expressed as:

$$Pr(T \le t;\boldsymbol{z},\boldsymbol{\theta}) = \pi(\boldsymbol{z};\boldsymbol{\theta}_1) + (1-\pi(\boldsymbol{z};\boldsymbol{\theta}_1))[1-S(t;\boldsymbol{\theta}_2|\boldsymbol{c}=0,\boldsymbol{z})] \quad (1)$$

- Stratified Weibull regression model

  To relax the proportional hazards assumption on the Weibull regression model, we use stratified cox model, where λ_z and γ_z will be depend on covariate z, and β is the covariate coefficient for confounder $\boldsymbol{x}$, so the survival model will be

$$S(t;\boldsymbol{\beta},\gamma_z,\lambda_z|\boldsymbol{x},\boldsymbol{z}) = \exp(-\lambda_{\boldsymbol{z}} t^{\gamma_z}\exp(\boldsymbol{x}^T\boldsymbol{\beta})) \quad (2)$$

- Likelihood

  Let $\boldsymbol{K}_1$ be subjects with $c_i=0$ or $c_i=1$. Let $\boldsymbol{K}_0$ be subjects whose c_i is missing at random (MAR).

So the observed log-likelihood ([1]) would be:

$$\begin{aligned}\ell(\boldsymbol{\theta}) = &\sum_{i\in\boldsymbol{K}_1}[c_i\log\{\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1)\} + (1-c_i)\log\{(1-\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1))\times \\ &(S(L_i;\boldsymbol{\theta}_2|c_i=0,z_i) - S(R_i;\boldsymbol{\theta}_2|c_i=0,z_i))\}] + \\ &\sum_{i\in\boldsymbol{K}_0}\log[\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1) + \{1-\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1)\}\{1-S(R_i;\boldsymbol{\theta}_2|c_i=0,z_i)\}]\end{aligned} \quad (3)$$

,where $\boldsymbol{\theta_1} = \{\alpha_0,\alpha_1\}$ and $\boldsymbol{\theta_2} = \{\beta,\lambda,\gamma\}$. $S(L_i;\boldsymbol{\theta}_2|c_i=0,z_i)$ and $S(R_i;\boldsymbol{\theta}_2|c_i=0,z_i))$ are the survival function at time L_i and R_i respectively.

Simulation

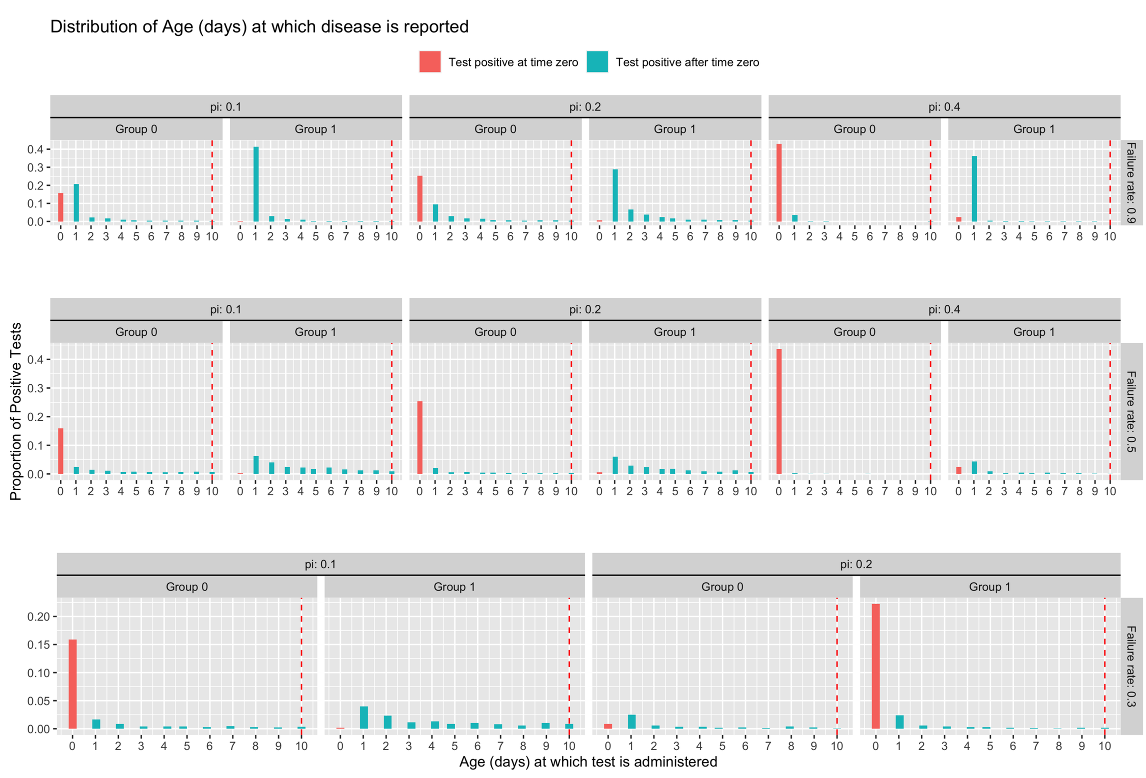


Fig. 1: In *Scenario I*, the probability of time to event at baseline $\pi(\theta_1|z)$ depend on covariate z and we assume PH assumption holds.

Fig. 2: In *Scenario II*, we relax the proportional hazards assumption in the Weibull regression model, so the baseline hazard depends on covariate z (Equation 2).

References

[1] Li C Cheung et al. “Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records”. In: *Statistics in medicine* 36.22 (2017), pp. 3583–3595.

[2] Xiangdong Gu et al. “Stratified Weibull regression model for interval-censored data”. In: *The R journal* 6.1 (2014), p. 31.

Simulation Results

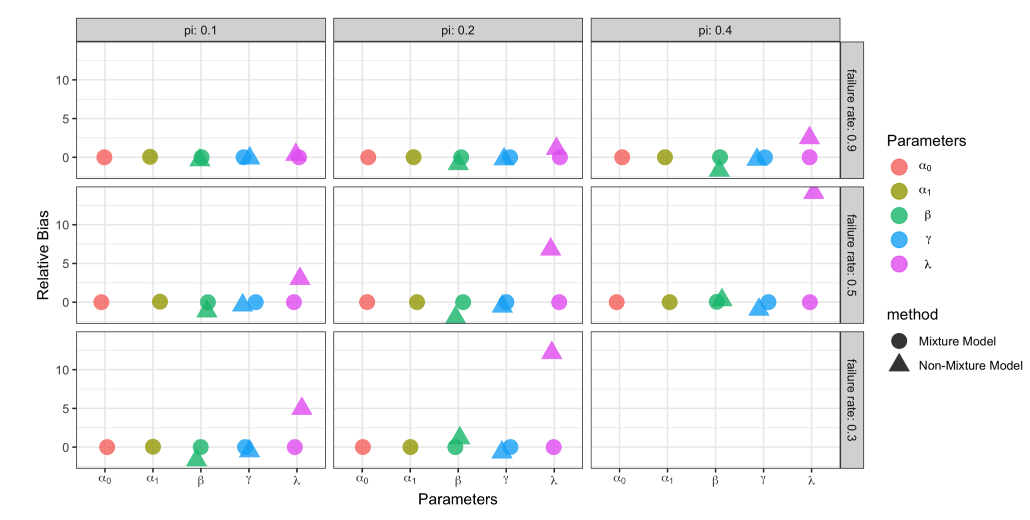


Fig. 3: Scenarios I: Relative Bias

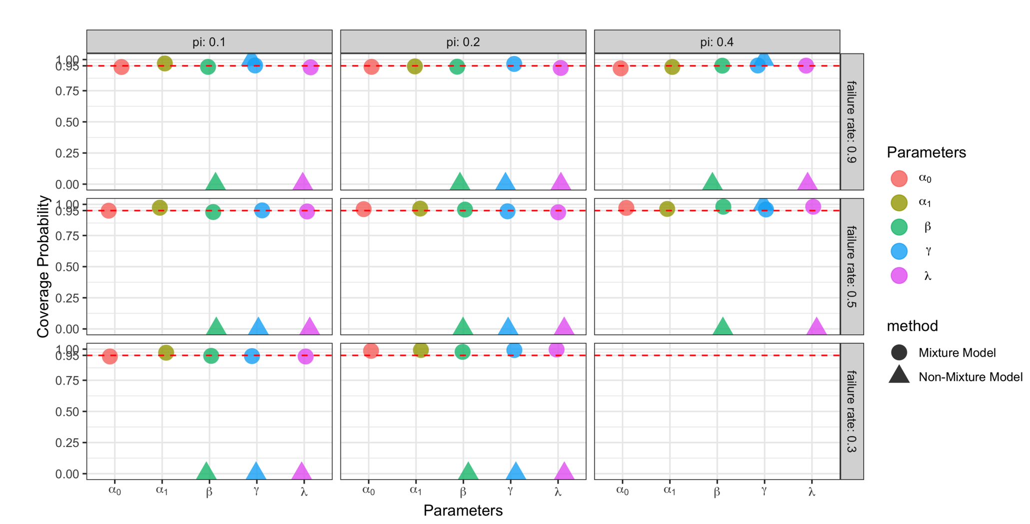


Fig. 4: Scenarios I: Coverage Probability

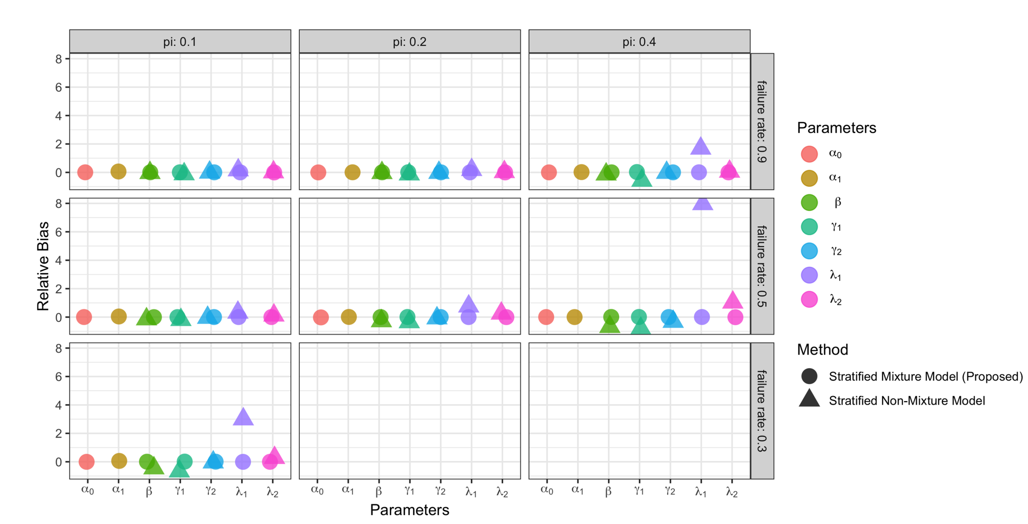


Fig. 5: Scenarios II: Relative Bias

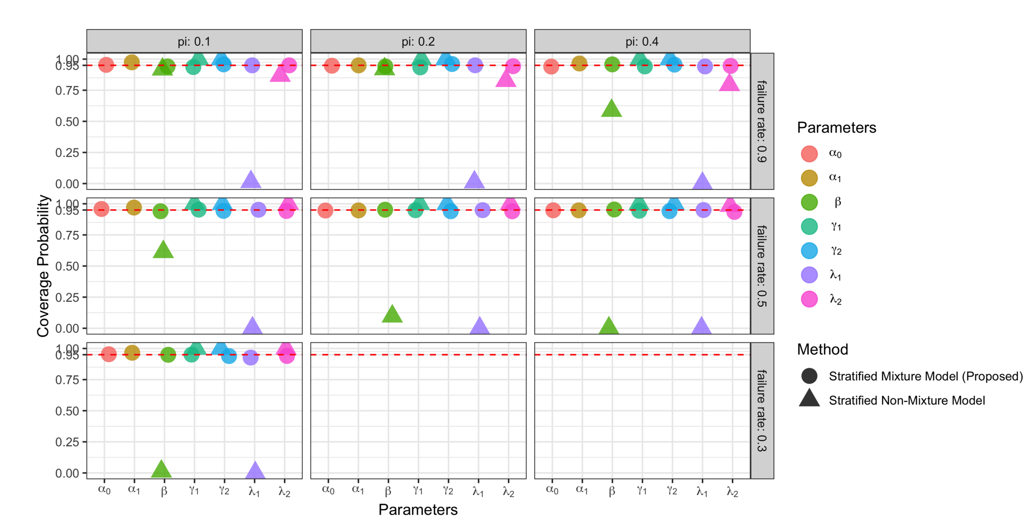


Fig. 6: Scenarios II: Coverage Probability

Application

- We apply proposed method to a subset from NHANES I (First National Health and Nutrition Examination Survey) Epidemiologic Follow-up Study (NHEFS). Data are available publicly via the CDC at https://wwwn.cdc.gov/nchs/nhanes/nhefs/default.aspx/.
- The goal is to
  - Estimate distribution of time to first report Diabetes in the interviews
  - Find potential risk factors for Diabetes

Check PH Assumption

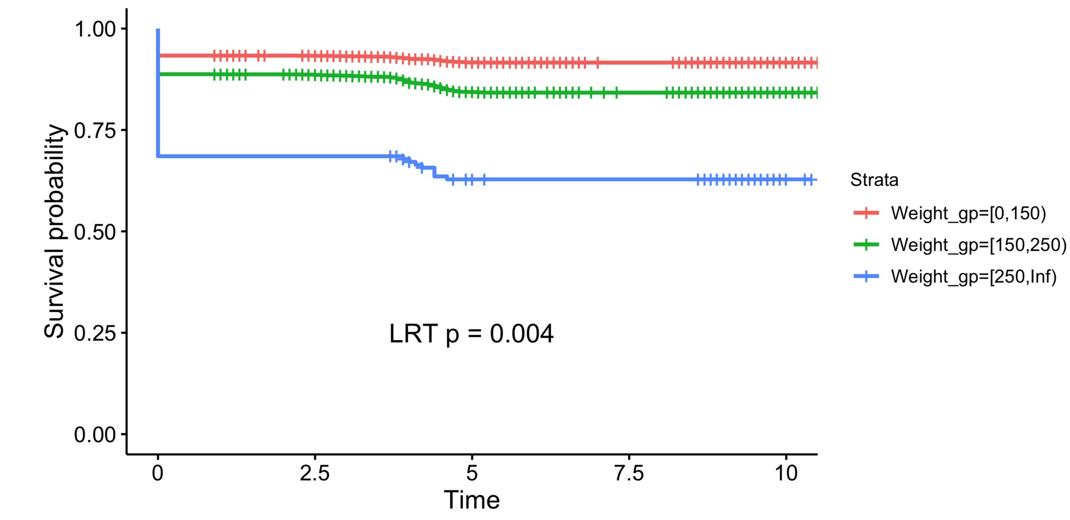


Fig. 7: We check the PH assumption by looking at survival probabilities in each weight group.

Application Results

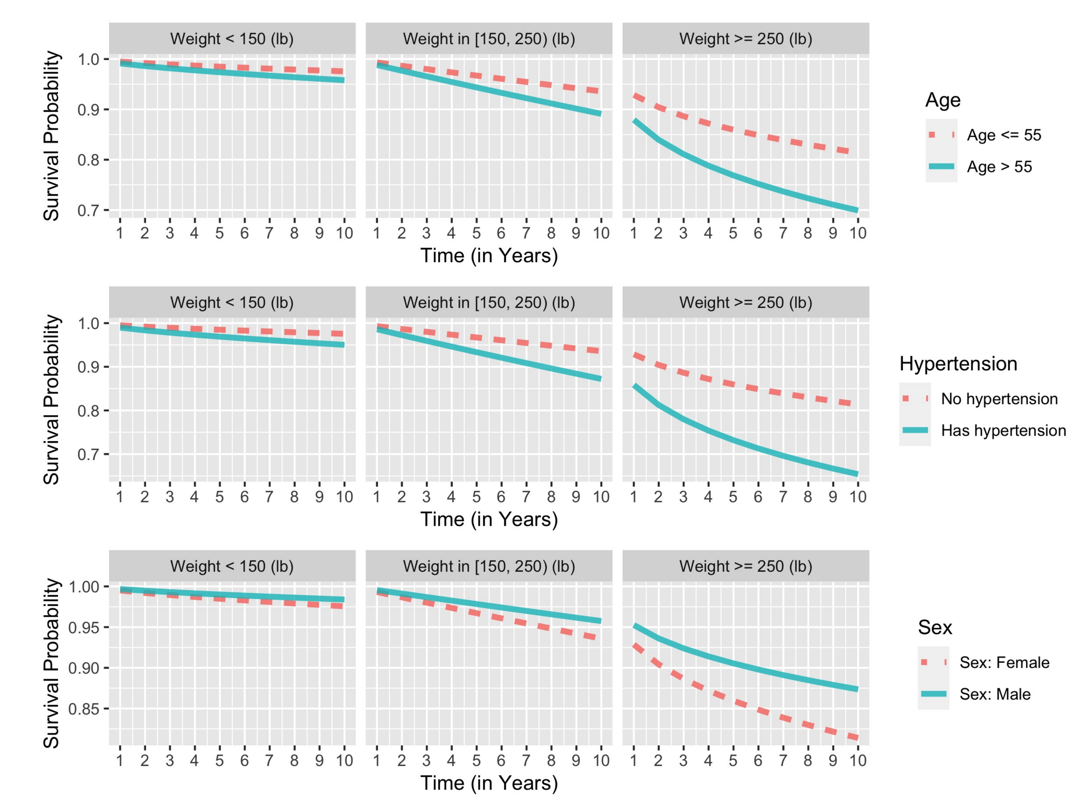


Fig. 8: Survival probability with respect to covariates stratified by weight groups.

Conclusion

- We used Logistic-Weibull mixture model to accommodate zero-inflated times, which performs better than non-mixture model, especially when proportion of positive tests at baseline is larger.
- When proportional hazards assumption does not hold, the proposed stratified Logistic-Weibull mixture model outperform non stratified model.
- In application, we used our proposed mixture stratified Weibull regression model to estimate the time to the diabetes. We found the risk factors for diabetes including weight ($\geq$250lb), age (> 55) and sex (Female).