Penalized regression for left-truncated & right-censored survival data

McGough, Incerti, Lyalina, Copping, Narasimhan, Tibshirani (2021) "Statistics in Medicine)

(A summary by Andrey Yeo Te-Jing).

Rationale: For some inclusion criterions for time to survival data, one sometimes include only patients with a genomic test in a personalised health care context (often the case for oncology medicine that this context is applied). plus with electronic health records (aka, just lots of variables) come high dimensional data (because they are)

(Problem → Solution proposed)

High dimensionality → Penalised Regression such as lasso (l_1) and ridge (l_2), elastic net regression, smoothly clipped absolute deviations

Left truncation at time t → Partial likelihood at time t : $\prod_{i}^{m} \frac{e^{x^T B}}{\sum e^{x^T B}}$

Overfitting → cross validation

andreytyeo@gmail.com

Things to avoid → Preventative Measure in Analysis

"immortal time bias" because observed patients cannot die prior to entering the study. → (that's a hard one to avoid when "letting" patients into a survival analysis study)

poor prediction accuracy + limited generalizability → Don't over fit add regularization penalty to constrain the size of B, reduce complexity of model

avoid p predictors > n sample (p<n is desirable)

audreytyeo@gmail.com

Simulation study (methods)

1) Data generation of) from weibull.

T_i = latent survival time

U_i = Right censoring

V_i = study entry time

} assumes they are uncorrelated

observed survival time = $Y_i = \min(T_i, U_i)$

R censor = $T_i > U_i$

L truncated = $V_i > Y_i$

2) Simulation of 10 binary predictors to an $n \times p$ matrix called X. (11 predictors are taken from real-world dataset already, s.t. $p=21$)

$$\begin{pmatrix} \tilde{X}^*_{i,1} \\ \vdots \\ \tilde{X}^*_{i,p} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \tilde{\mu_p} \end{pmatrix}, \Sigma \right)$$

} multivariate probit approach

T_i, latent survival is taken from a Weibull survival function with density

$$f(t) = a m t^{a-1} e^{mta}$$

where $m = \exp(\alpha + x^T B)$

s.t. $T_i \sim \text{Weibull}(\alpha, m_i)$

choice because when using Nelson-Aalen estimator of $\Phi(t)$, cumulative hazard, overall survival is monotonically decreasing, so Weibull can capture baseline hazard

audreytyeo@gmail.com

U_i, right censoring time is taken also from a Weibull distribution

V_i, study entry times is generated in 2 parts, first is the bernoulli random variable $d_i \sim$ Bernoulli($\tilde{\pi}$) to input 4 V has status 1 or 0.

$V \sim$ lognormal (μ, σ^2), median = 1 year, mean = 1.6 years

not sure rational for choice maybe convenience from Normal distribution as inverse cdf and pdf can be "easily" computed and (R) skewed (see mean, median chosen)

3) cross validation with 10 folds where test data is 25% and training is 75% from ~4500 observations of real-world data
repeated 200 times

3a)
3b) from small and large p scenarios

$p=21$ $p=1011$

4) find λ that minimized partial likelihood deviance in (3)

5) plot calibration curves of survival probability of those with and without calibration curves. for small and large p scenario

audreytyeo@gmail.com

pg 4 of 6

6) Use C-index to compare fitted models with and without Left truncation adjustment. the ↑ C-index, the better the model is to discriminate prognosis (survival probability) between patients

7) with penalised Cox proportional model, compare hazard ratios between models that were and were not adjusted for Left truncation

Results (1 example)

model	Covariate size	C-index adjustment	 no adjustment
Cox	small	0.649	0.580
Cox (lasso)	small	0.648 not better	0.580 no diff
Cox	large	0.625	0.556
Cox (lasso)	large	0.663 better	0.600 better

andreytyco@gmail.com

Points of Discussion

→ generally C-index is higher in model that does not adjust for Left truncation ? maybe because of bias in test = training sets. therefore careful understanding of the data generation process is crucial ⚠

→ this approach opens questions for future specification, interpretation and evaluation of prognostic survival models

→ Calibration curves > C-index, maybe latter is more difficult to differentiate between models

→ define risk set appropriately, difficult wh Ⓛ truncated data ∴ when and how patients came to be included in the data

→ potential selection bias because only stage IV diagnoses was included eg temporal selection bias, immortal time bias. Can be mitigated by studying the association between left truncated time and survival time.

audreytyeo@gmail.com