Simulation Study of a MLE and Bootstrap-based Goodness-of-fit Test for Parametric Generalized Linear Models under Random Censorship

Gitte Kremling[1,2], Gerhard Dikta[1] and Richard Stockbridge[2]

[1]Fachhochschule Aachen, Fachbereich 9, Medizintechnik und Technomathematik, Heinrich-Mußmann-Str. 1, 52428 Jülich, Germany, E-Mail: kremling@fh-aachen.de

[2]University of Wisconsin-Milwaukee, Department of Mathematical Sciences, PO Box 413, Milwaukee, WI 53201-0413, USA

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

What is the problem? – Data and model

Given censored survival data with corresponding covariates, check whether data fits to parametric generalized linear model (GLM).

Underlying data:

- X – covariates in $\mathbb{R}^p$
- Y – survival time
- C – censoring time

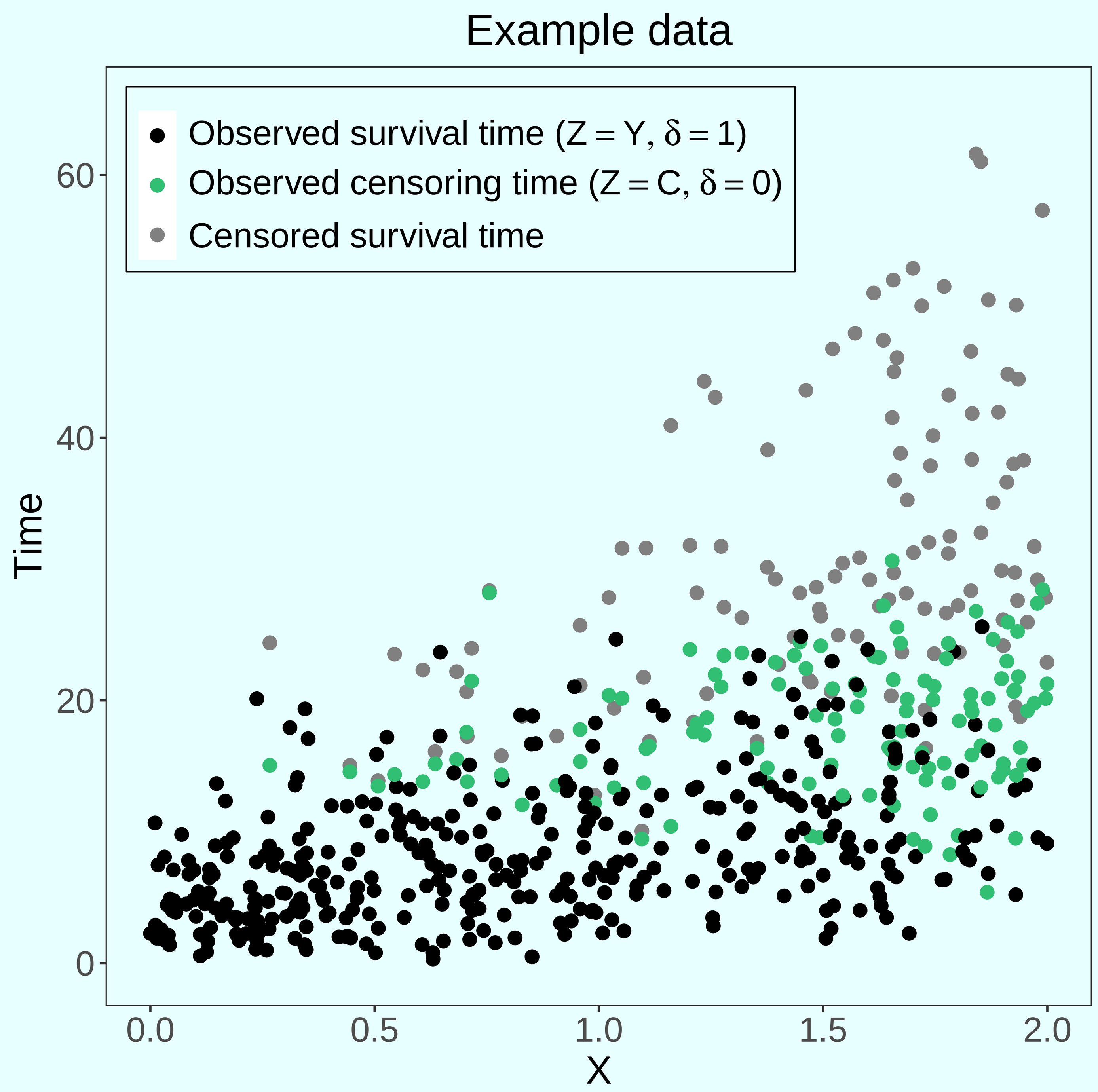
Observed data:

- X – covariates in $\mathbb{R}^p$
- Z – event time $Z = \min\{Y, C\}$
- δ – censoring indicator $\delta = \mathbb{I}_{\{Y \leq C\}}$

Parametric generalized linear model:

(i) Distribution $F_{Y|X}$ of Y given X belongs to a given exponential family with dispersion parameter ϕ

(ii) $g(\mathbb{E}[Y|X = x]) = \beta^T x$ for some $\beta \in \mathbb{R}^p$ and a given link function g



How do we tackle it? – Goodness-of-fit test

1. Compute **MLE** for $\hat{\beta}_n$ and $\hat{\phi}_n$
2. Use difference between parametric and non-parametric fit for distribution of Y as **test statistic**:
$$\tilde{\alpha}_n^{KM}(t) = \sqrt{n}\left(F_n^{KM}(t) - \hat{F}_n(t|\hat{\beta}_n, \hat{\phi}_n)\right)$$
with Kaplan-Meier estimate $F_n^{KM}(t) = 1 - S_n^{KM}(t)$ and $\hat{F}_n(t|\hat{\beta}_n, \hat{\phi}_n) = \frac{1}{n}\sum_{i=1}^{n} F_{Y|X}(t|X_i; \hat{\beta}_n, \hat{\phi}_n)$
3. Compute e.g. Kolmogorov-Smirnov type distance $D_n := \sup_t |\tilde{\alpha}_n^{KM}(t)|$
4. Estimate the p-value of D_n using a **parametric bootstrap**

Let's see how well it works! – Simulation study

H_0	$Y\|X \sim \text{Gamma}(\phi)$	$\log(\mathbb{E}[Y\|X = x]) = \beta^T x$
Sim. (A)	$Y\|X \sim \text{Gamma}, \phi = 1$	$\log(\mathbb{E}[Y\|X = x]) = x_1 + 2x_2$
Sim. (B)	$Y\|X \sim \text{Gamma}, \phi = 1$	$\log(\mathbb{E}[Y\|X = x]) = x_1 + 2x_2 + 0.1x_2^2$
Sim. (C)	$Y\|X \sim \text{Normal}, \phi = 1$	$\log(\mathbb{E}[Y\|X = x]) = x_1 + 2x_2$

$X_1 = 1,\ X_2 \sim UNI(-5, 5)$,
$C \sim \mathcal{N}(9, 1)$ ($\approx 40\%$ censored)

$n = 500$ observations
$m = 100$ bootstrap iterations
$rep = 100$ simulation repetitions

Distribution of MLE under (A)

	β_1	β_2	ϕ
TrueValues	1	2	1
Mean	0.9976	2.0007	0.9835
Variance	0.0080	0.0011	0.0055
MSE	0.0080	0.0011	0.0058