FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Bootstrap-based goodness-of-fit test for parametric generalized linear models under random censorship

July 6, 2023

Gitte Kremling, Gerhard Dikta, Richard Stockbridge

UNIVERSITY of WISCONSIN
UWMILWAUKEE

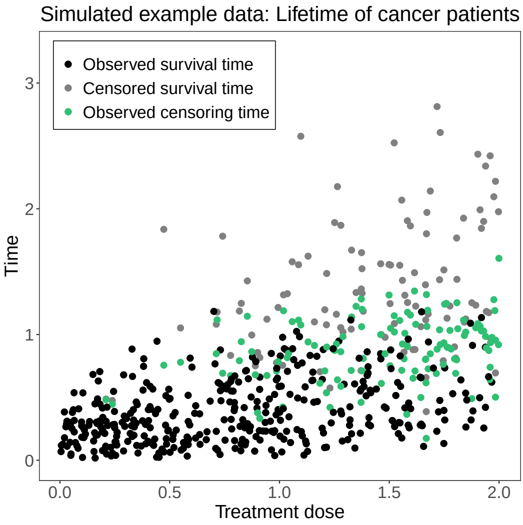
Motivating Example - Survival Analysis

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

- Medical study about the lifetime of cancer patients after treatment starts
- Data is randomly right-censored (study ends / patients drop out)
- Covariates such as treatment dose or age of patient are fully observed

Interested in distribution of survival times dependent on covariates.

Motivating Example - Survival Analysis



Mathematical Framework

Underlying data:

- covariates $X_i \in \mathbb{R}^p$
- survival times $Y_i \in \mathbb{R}_+$
- censoring times $C_i \in \mathbb{R}_+$

Observed data:

- covariates $X_i \in \mathbb{R}^p$
- censored times $Z_i = \min(Y_i, C_i)$
- censoring indicators $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$

Mathematical Framework

Underlying data:

- covariates $X_i \in \mathbb{R}^p$
- survival times $Y_i \in \mathbb{R}_+$
- censoring times $C_i \in \mathbb{R}_+$

Observed data:

- covariates $X_i \in \mathbb{R}^p$
- censored times $Z_i = \min(Y_i, C_i)$
- censoring indicators $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$

Problem: Given an i.i.d. sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$, find the distribution of survival times Y dependent on the vector of covariates X

Here: Check whether data fits to a parametric generalized linear model.

[Assumption: C is independent of $\sigma(X, Y)$]

Parametric Generalized Linear Model (GLM)

- Linear Model:
  - $\mathbb{E}[Y|X = x] = \beta^T x$ for some $\beta \in \mathbb{R}^p$

Parametric Generalized Linear Model (GLM)

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

Parametric Generalized Linear Model (GLM)

- Parametric Generalized Linear Model:
  - $g(\mathbb{E}[Y|X = x]) = \beta^T x$ for some $\beta \in \mathbb{R}^p$ and some given link function g
  - $F_{Y|X}$ belongs to an exponential family with dispersion parameter ϕ

Parametric Generalized Linear Model (GLM)

- Parametric Generalized Linear Model:
  - $g(\mathbb{E}[Y|X = x]) = \beta^T x$ for some $\beta \in \mathbb{R}^p$
    and some given link function g
  - $F_{Y|X}$ belongs to an exponential family
    with dispersion parameter ϕ
- These two hypotheses can be combined into a single one:

$$H_0 : Y|X \sim F_{Y|X} \in \{F(\,\cdot\,|X, \beta, \phi)|\beta \in \mathbb{R}^p, \phi > 0\}$$

Goodness-of-fit test - Test statistic

- Difference between parametric and non-parametric estimate of marginal distribution function F_Y
- Parametric fit, using MLE $(\hat{\beta}_n, \hat{\phi}_n)$ and ecdf $\hat{H}_n$ of $\{X_i\}_{i=1}^n$:

$$\hat{F}_Y(t|\hat{\beta}_n, \hat{\phi}_n) = \int F(t|x, \hat{\beta}_n, \hat{\phi}_n)\hat{H}_n(dx) = \frac{1}{n}\sum_{i=1}^{n} F(t|X_i, \hat{\beta}_n, \hat{\phi}_n)$$

- Non-parametric fit: Kaplan-Meier estimator

Goodness-of-fit test - Test statistic

- Difference between parametric and non-parametric estimate of marginal distribution function F_Y
- Parametric fit, using MLE $(\hat{\beta}_n, \hat{\phi}_n)$ and ecdf $\hat{H}_n$ of $\{X_i\}_{i=1}^n$:

$$\hat{F}_Y(t|\hat{\beta}_n, \hat{\phi}_n) = \int F(t|x, \hat{\beta}_n, \hat{\phi}_n)\hat{H}_n(dx) = \frac{1}{n}\sum_{i=1}^{n} F(t|X_i, \hat{\beta}_n, \hat{\phi}_n)$$

- Non-parametric fit: Kaplan-Meier estimator
- Kaplan-Meier type empirical process with estimated parameters and covariates:

$$\tilde{\alpha}_n^{\mathsf{KM}}(t) = \sqrt{n}\left(\hat{F}_{Y,n}^{\mathsf{KM}}(t) - \hat{F}_Y(t|\hat{\beta}_n, \hat{\phi}_n)\right)$$

- Use e.g. Kolmogorov-Smirnov distance $\|\tilde{\alpha}_n^{\mathsf{KM}}\| = \sup_t |\tilde{\alpha}_n^{\mathsf{KM}}(t)|$

Goodness-of-fit test - Test statistic

- Difference between parametric and non-parametric estimate of marginal distribution function F_Y
- Parametric fit, using MLE $(\hat{\beta}_n, \hat{\phi}_n)$ and ecdf $\hat{H}_n$ of $\{X_i\}_{i=1}^n$:

$$\hat{F}_Y(t|\hat{\beta}_n, \hat{\phi}_n) = \int F(t|x, \hat{\beta}_n, \hat{\phi}_n)\hat{H}_n(dx) = \frac{1}{n}\sum_{i=1}^{n} F(t|X_i, \hat{\beta}_n, \hat{\phi}_n)$$

- Non-parametric fit: Kaplan-Meier estimator
- Kaplan-Meier type empirical process with estimated parameters and covariates:

$$\tilde{\alpha}_n^{\mathsf{KM}}(t) = \sqrt{n}\left(\hat{F}_{Y,n}^{\mathsf{KM}}(t) - \hat{F}_Y(t|\hat{\beta}_n, \hat{\phi}_n)\right)$$

- Use e.g. Kolmogorov-Smirnov distance $\|\tilde{\alpha}_n^{\mathsf{KM}}\| = \sup_t |\tilde{\alpha}_n^{\mathsf{KM}}(t)|$

Find distribution of $\|\tilde{\alpha}_n^{\mathsf{KM}}\|$ to decide when H_0 should be rejected.

Goodness-of-fit test - Limit distribution

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Theorem

Under H_0 and some regularity conditions, $\tilde{\alpha}_n^{KM}$ converges in $D[0,T]$ to a centered Gaussian process $\tilde{\alpha}_\infty^{KM}$ with known covariance function which is dependent on the true distribution functions of X, Y and C.

Goodness-of-fit test - Limit distribution

Theorem

Under H_0 *and some regularity conditions,* $\tilde{\alpha}_n^{KM}$ *converges in* $D[0,T]$ *to a centered Gaussian process* $\tilde{\alpha}_\infty^{KM}$ *with known covariance function which is dependent on the true distribution functions of* X, Y *and* C*.*

Proof sketch:

- Splitting as in Durbin (1973):

$$\begin{aligned}\tilde{\alpha}_n^{\mathsf{KM}}(t) &= \sqrt{n}\Big(\hat{F}_{Y,n}^{\mathsf{KM}}(t) - F_Y(t)\Big) \\ &\quad + \sqrt{n}\left(F_Y(t) - \hat{F}_Y(t|\beta_0,\phi_0)\right) \\ &\quad + \sqrt{n}\left(\hat{F}_Y(t|\beta_0,\phi_0) - \hat{F}_Y(t|\hat{\beta}_n,\hat{\phi}_n)\right)\end{aligned}$$

- Apply Billingsley (1968), Theorem 15.1:
  convergence of fidis $\wedge$ tightness $\Rightarrow$ convergence in $D[0,T]$

Goodness-of-fit test - Limit distribution

Theorem

Under H_0 and some regularity conditions, $\tilde{\alpha}_n^{KM}$ converges in $D[0,T]$ to a centered Gaussian process $\tilde{\alpha}_\infty^{KM}$ with known covariance function which is dependent on the true distribution functions of X, Y and C.

$\boxed{\text{Use bootstrap to approximate the distribution of } \|\tilde{\alpha}_n^{\mathsf{KM}}\|.}$

Goodness-of-fit test - Bootstrap

Goal: Estimate the distribution of $\|\tilde{\alpha}_n^{KM}\|$ under H_0

Idea:

- Resample from the given data in a way that H_0 is fulfilled $\left(Y_i^* \sim F(\,\cdot\,|X_i^*, \hat{\beta}_n, \hat{\phi}_n)\right)$
- Compute the test statistic $\|\tilde{\alpha}_n^{KM*}\|$ for this new sample
- Repeat these steps many times and use the empirical distribution function of the resulting test statistics as an estimate

Result: p-value is given by the percentage of $\|\tilde{\alpha}_n^{KM*}\|$ that are greater than or equal to $\|\tilde{\alpha}_n^{KM}\|$

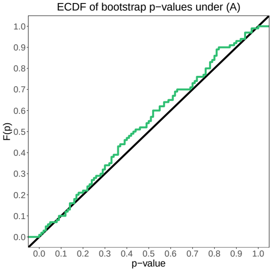
Numerical results - Simulated data

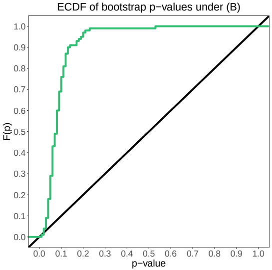
H_0	$Y\|X \sim$ Gamma(ϕ)	$\log(\mathbb{E}[Y\|X=x]) = \beta^T x$
Sim. (A)	$Y\|X \sim$ Gamma, $\phi = 1$	$\log(\mathbb{E}[Y\|X=x]) = x_1 + 2x_2$
Sim. (B)	$Y\|X \sim$ Gamma, $\phi = 1$	$\log(\mathbb{E}[Y\|X=x]) = x_1 + 2x_2 + 0.1x_2^2$
Sim. (C)	$Y\|X \sim$ Normal, $\phi = 1$	$\log(\mathbb{E}[Y\|X=x]) =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

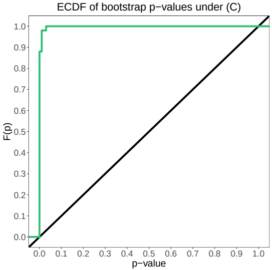
- Covariates $X = (X_1, X_2)$ with $X_1 = 1$, $X_2 \sim UNI(-5, 5)$
- Censoring times $C \sim \mathcal{N}(9, 1)$ ($\approx 40\%$ censored)
- $n = 500$ observations, $m = 100$ bootstrap iterations, $rep = 100$ simulation repetitions

Numerical results - Goodness-of-fit test

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







Conclusions & Outlook

- Developed a goodness-of-fit test for parametric GLM under random censorship
- Promising numerical results
- Identified the limit distribution of the test statistic (Kaplan-Meier type process with estimated parameters and covariates)

Conclusions & Outlook

- Developed a goodness-of-fit test for parametric GLM under random censorship
- Promising numerical results
- Identified the limit distribution of the test statistic (Kaplan-Meier type process with estimated parameters and covariates)

Next:

- Identify the limit distribution of the corresponding bootstrap process (should be the same)
- Apply methods to a real data example

References I

Billingsley, Patrick (1968). *Convergence of probability measures*. John Wiley & Sons Inc., New York.

Durbin, James (1973). "Weak convergence of the sample distribution function when parameters are estimated". In: *The Annals of Statistics*, pp. 279–290.

Kaplan, Edward L and Paul Meier (1958). "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282, pp. 457–481.

Nikabadze, A and W Stute (1997). "Model checks under random censorship". In: *Statistics & probability letters* 32.3, pp. 249–259.

Stute, W, W González Manteiga, and C Sánchez Sellero (2000). "Nonparametric model checks in censored regression". In: *Communications in Statistics-theory and Methods* 29.7, pp. 1611–1629.

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Conclusions & Outlook

- Developed a goodness-of-fit test for parametric GLM under random censorship
- Promising numerical results
- Identified the limit distribution of the test statistic (Kaplan-Meier type process with estimated parameters and covariates)

Next:

- Identify the limit distribution of the corresponding bootstrap process (should be the same)
- Apply methods to a real data example

Thank you for your attention! Any questions? :)

Kaplan-Meier estimator

- Non-parametric estimator of the survival function of Y

$$S_Y(t) = 1 - F_Y(t) = \mathbb{P}(Y > t)$$

given a censored i.i.d. sample $(Z_i, \delta_i)_{i=1}^n$

- If S_Y is discrete with mass at points $t_1 < ... < t_n$,

$$S_Y(t) = \prod_{i:t_i \leq t} \mathbb{P}(Y > t_i | Y \geq t_i) = \prod_{i:t_i \leq t} (1 - \mathbb{P}(Y = t_i | Y \geq t_i))$$

- Kaplan-Meier (KM) estimator defined by

$$\hat{S}_{Y,n}^{KM}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

t_i: time when at least one event happened
$d_i = \sum_{i=1}^n \delta_i \mathbb{1}_{\{Z_i = t_i\}}$ (number of events that happened at time t_i)
$n_i = \sum_{i=1}^n \mathbb{1}_{\{Z_i \geq t_i\}}$ (individuals known to have survived up to time t_i)

Goodness-of-fit test - Resampling scheme

1. For $i = 1, \ldots, n$
   a) Generate X_i^* according to the empirical df of $X_1, \ldots, X_n$
   b) Generate Y_i^* according to the parametric fit $F_{Y|X}(\,\cdot\,|X_i^*, \hat{\beta}_n, \hat{\phi}_n)$
   c) Generate C_i^* according to the Kaplan-Meier estimator for the censoring times $C_1, \ldots, C_n$
   d) Set $Z_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = \mathbb{1}_{\{Y_i^* \leq C_i^*\}}$
2. Compute MLE ($\hat{\beta}_n^*$, $\hat{\phi}_n^*$) for bootstrap data set $(X_i^*, Z_i^*, \delta_i^*)_{i=1}^n$
3. Obtain process $\tilde{\alpha}_n^{\mathsf{KM}*}(t)$ and calculate KS/CvM distance
4. Repeat steps 1-3 m times to compute bootstrap p-value