Censored count data regression with missing censoring information

J.-F. Dupuy (IRMAR Rennes)
avec B. Bousselmi (IRMAR Rennes) et A. Karoui (Univ. Carthage)

ASMSA 2020
Poitiers, 10-11/12/2020

Outline

1. Poisson regression model and the problem
2. The multiple imputation estimator
3. A simulation study
4. Concluding remarks

1 Poisson regression model and the problem

2 The multiple imputation estimator

3 A simulation study

4 Concluding remarks

Poisson regression model : setup

- a model for count data, e.g. numbers of insurance claims, of cases of some disease
- for homogeneous data $\{Y_1, \ldots, Y_n\}$ with $Y_i \sim \mathcal{P}(\lambda)$,

$$\mathbb{P}(Y_i = y) = e^{-\lambda} \frac{\lambda^y}{y!}, \qquad \text{for all } y \in \mathbb{N}$$

- regression setting : $\{(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)\}$ with predictors $\mathbf{X}_i = (1, X_{2i}, \ldots, X_{pi})^\top$ and

  $Y_i | \mathbf{X}_i \sim \mathcal{P}(\lambda_i)$ with $\lambda_i = \exp(\beta^\top \mathbf{X}_i)$ (log-linear model)

  $\hookrightarrow$ Poisson regression is a particular case of generalized linear model (GLM)

Poisson regression model : inference

Based on $\{(y_1, \mathbf{X}_1), \ldots, (y_n, \mathbf{X}_n)\}$, the log-likelihood of β in the model $Y_i|\mathbf{X}_i \sim \mathcal{P}(\exp(\beta^\top \mathbf{X}_i))$ is :

$$\begin{aligned} \ell_n(\beta) &= \log\left\{\prod_{i=1}^n \mathbb{P}(Y_i = y_i|\mathbf{X}_i)\right\} \\ &= \sum_{i=1}^n \left\{y_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(y_i!)\right\}, \end{aligned}$$

and if $\hat{\beta}_n = \arg\max_\beta \ell_n(\beta)$, then :

$$\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta \quad \text{and} \quad \sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\ell} \mathcal{N}(0, \Sigma)$$

where $\Sigma = (\mathbb{E}[\mathbf{X}\mathbf{X}^\top e^{\beta^\top \mathbf{X}}])^{-1}$ can be consistently estimated ($\hookrightarrow$ tests of hypothesis, confidence intervals).

Censored Poisson regression

The count Y_i is right-censored if the true count is higher than the observed one (e.g., smoking habits, healthcare utilization. . .)

Modeling :

- let Y_i^* be the observed count. We only know that one of these events has occurred :

$$\mathcal{E}_i = \{Y_i = Y_i^*\} \text{ or } \mathcal{F}_i = \{Y_i \geq Y_i^*\}$$

- let $\delta_i = 1_{\mathcal{E}_i}$ be the **censoring indicator**

Observations : $\{(Y_1^*, \delta_1, \mathbf{X}_1), \ldots, (Y_n^*, \delta_n, \mathbf{X}_n)\}$

- without censoring, we observe : $\{(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)\}$

Censored Poisson regression : inference

Maximum likelihood estimation (MLE) :

$$\prod_{i=1}^{n} \mathbb{P}(Y_i = y_i^* | \mathbf{X}_i)^{\delta_i} \mathbb{P}(Y_i \geq y_i^* | \mathbf{X}_i)^{1-\delta_i}$$

from which we easily deduce the loglikelihood :

$$\sum_{i=1}^{n} \delta_i \left(y_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(y_i^*!) \right)$$
$$+ \sum_{i=1}^{n} (1-\delta_i) \log \left(1 - \sum_{k=0}^{Y_i^*-1} \frac{e^{-\exp(\beta^\top \mathbf{X}_i) + k\beta^\top \mathbf{X}_i}}{k!} \right)$$

Standard asymptotic theory applies to the censored MLE.

In the literature : Terza (1985), Famoye and Wang (2004), Karlis et al. (2016), Nguyen and Dupuy (2020). . .

A problem of missing data

- the censoring indicator δ_i may be missing for some individuals
- let ξ_i be the missingness indicator, that is

$$\xi_i = 1 \text{ if } \delta_i \text{ is observed and } \xi_i = 0 \text{ otherwise}$$

- observed data for individual i are :

$$(Y_i^*, \mathbf{X}_i, \delta_i, \xi_i = 1) \quad \text{or} \quad (Y_i^*, \mathbf{X}_i, \xi_i = 0).$$

- missing-at-random (MAR) mechanism : $\xi \perp\!\!\!\perp \delta$ given all other observed variables (less restrictive than MCAR)

In the literature : survival analysis : Subramanian (2006, 2011), Wang et al. (2012), Brunel et al. (2013)... linear regression : Wang and Dinse (2011)

1 Poisson regression model and the problem

2 The multiple imputation estimator

3 A simulation study

4 Concluding remarks

Multiple imputation (MI)

Basic idea :

1. generate M imputed (completed) datasets $\hookrightarrow$ imputation model
2. fit the model on each imputed dataset
3. combine inferences from the M imputed datasets (Rubin's rules) :
   - pool the M estimates into a single MI estimate
   - pool the M variance estimates

Imputation model : a model for the conditional distribution of δ_i given the observed data $\mathbf{W}_i = (Y_i^*, \mathbf{X}_i)$. Here

$$\delta_i | \mathbf{W}_i \sim \mathcal{B}(\mathbb{E}(\delta_i | \mathbf{W}_i))$$

Assume a model for the unknown $\mathbb{E}(\delta_i | \mathbf{W}_i)$. Several possibilities :

Imputation model

- parametric regression model :

$$G(\mathbb{E}(\delta_i|\mathbf{W}_i)) = \theta^\top \mathbf{W}_i, \quad \theta \in \mathbb{R}^k$$

with $G(\cdot)$ a known *link* function, e.g. probit, logit :

$$\mathbb{E}(\delta_i|\mathbf{W}_i) = G^{-1}(\theta^\top \mathbf{W}_i) = \frac{e^{\theta^\top \mathbf{W}_i}}{1 + e^{\theta^\top \mathbf{W}_i}}$$

(may include polynomial terms, interactions. . .)

- semiparametric alternative : the single-index model

$$G(\mathbb{E}(\delta_i|\mathbf{W}_i)) = h(\theta^\top \mathbf{W}_i), \quad \theta \in \mathbb{R}^k$$

with $h(\cdot)$ an unknown function. E.g.

$$\mathbb{E}(\delta_i|\mathbf{W}_i) = \frac{e^{h(\theta^\top \mathbf{W}_i)}}{1 + e^{h(\theta^\top \mathbf{W}_i)}}$$

Imputation model

- semiparametric alternative : the generalized additive model (GAM)

$$G(\mathbb{E}(\delta_i|\mathbf{W}_i)) = \theta + h_2(W_{2i}) + \ldots + h_K(W_{Ki}), \quad \theta \in \mathbb{R}$$

  where $h_2, \ldots, h_K$ are unknown functions of the components of $\mathbf{W}_i = (1, W_{2i}, \ldots, W_{Ki})$

- nonparametric alternative :

$$G(\mathbb{E}(\delta_i|\mathbf{W}_i)) = h(\mathbf{W}_i)$$

  with $h(\cdot)$ some unknown function

- many others...

Imputation model

- In this work, we assume a parametric model

$$\mathbb{E}(\delta_i|\mathbf{W}_i) = m(\mathbf{W}_i, \theta), \quad \theta \in \mathbb{R}^k$$

- estimate θ by MLE on complete cases $\{i = 1, \ldots, n | \xi_i = 1\}$:

$$\hat{\theta}_n = \arg\max_{\theta} \prod_{i=1}^{n} m(\mathbf{W}_i, \theta)^{\xi_i \delta_i} (1 - m(\mathbf{W}_i, \theta))^{\xi_i (1-\delta_i)}$$

- by standard MLE theory, $\hat{\theta}_n$ verifies :

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Theta^{-1}(\theta) \widetilde{m_i}(\theta) \xi_i (\delta_i - m(\mathbf{W}_i, \theta)) + o_{\mathbb{P}}(1)$$

Multiple imputation (ctd)

1. each missing δ_i is replaced by a random draw

$$D_{i,j}(\hat{\theta}_n) \sim \mathcal{B}(m(\mathbf{W}_i, \hat{\theta}_n)), \quad j = 1, \ldots, M$$

2. for each dataset j, let $\delta^*_{i,j}(\hat{\theta}_n) = \xi_i \delta_i + (1 - \xi_i) D_{i,j}(\hat{\theta}_n)$ and

$$\hat{\beta}^*_{n,j} = \arg\max_{\beta} \left\{ \sum_{i=1}^{n} \delta^*_{i,j}(\hat{\theta}_n) \left(y_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(y_i^*!) \right) \right.$$
$$\left. + \sum_{i=1}^{n} (1 - \delta^*_{i,j}(\hat{\theta}_n)) \log \left(1 - \frac{\Gamma\left(y_i^*, e^{\beta^\top \mathbf{X}_i}\right)}{(y_i^* - 1)!} \right) \right\}$$

where

$$\mathbb{P}(\mathcal{P}(\lambda) \leq u) = \frac{\Gamma(u+1, \lambda)}{u!} \text{ and } \Gamma(u, \lambda) = \int_{\lambda}^{\infty} t^{u-1} \exp(-t) dt$$

Multiple imputation (ctd)

3. average the M estimators $\hat{\beta}^*_{n,j}, j = 1, \ldots, M$ as :

$$\hat{\beta}^*_n = \frac{1}{M} \sum_{j=1}^{M} \hat{\beta}^*_{n,j}.$$

4. Rubin's rule for variance estimation in MI :
   - within-imputation variance : $W = \frac{1}{M} \sum_{j=1}^{M} \widehat{\text{var}\left(\hat{\beta}^*_{n,j}\right)}$
   - between-imputation variance : $B = \frac{1}{M-1} \sum_{j=1}^{M} (\hat{\beta}^*_{n,j} - \hat{\beta}^*_n)^2$

Final variance estimate for $\hat{\beta}^*_n$: $F = W + \left(1 + M^{-1}\right) B$

$\hookrightarrow$ alternative : derive asymptotic variance of the MI estimate $\hat{\beta}^*_n$

Asymptotics for the MI estimator

Theorem 1 (BDK, 2020)

Under some regularity conditions, as $n \longrightarrow \infty$, $\hat{\beta}_n^* \xrightarrow{\mathbb{P}} \beta$ *and* $\sqrt{n}(\hat{\beta}_n^* - \beta) \xrightarrow{\ell} \mathcal{N}(0, \Sigma^*)$, *where*

$$\Sigma^* = \Sigma_1^{-1}(\beta)\left\{\Sigma_4(\beta,\theta) + (2\Sigma_3(\beta,\theta) - \Sigma_2(\beta,\theta))\right.$$
$$\left.\Theta^{-1}(\theta)\Sigma_2^\top(\beta,\theta)\right\}\Sigma_1^{-1}(\beta),$$

for some matrices $\Sigma_1, \Sigma_2, \Sigma_3$ *and* Σ_4. *Moreover, there exists a consistent estimator of* Σ^*.

Remark 1

With no missing data, Σ^* reduces to $\Sigma_1^{-1}(\beta)$ (asymptotic variance of the MLE in censored Poisson regression).

A note on regression calibration

Alternative approach to missing data : missing data are replaced by their conditional expectation given observed data.

- multiple imputation :

$$\delta^*_{i,j}(\hat{\theta}_n) = \xi_i \delta_i + (1-\xi_i) D_{i,j}(\hat{\theta}_n) \in \{0,1\}, \quad j = 1, \ldots, M$$

  where $D_{i,j}(\hat{\theta}_n) \sim \mathcal{B}(m(\mathbf{W}_i, \hat{\theta}_n))$

- regression calibration :

$$\hat{\delta}_i(\hat{\theta}_n) = \xi_i \delta_i + (1-\xi_i) m(\mathbf{W}_i, \hat{\theta}_n) \in [0,1]$$

Both require a model for $\mathbb{E}(\delta_i | \mathbf{W}_i) \longrightarrow$ misspecification issue

A note on inverse probability weighting (IPW)

A different philosophy :

- use complete cases $\{i = 1, \ldots, n | \xi_i = 1\}$ only but :
- weights individuals contributions (in the likelihood) by the inverse of the selection probability, that is :

$$\frac{1}{\mathbb{P}(\xi_i = 1 | \mathbf{W}_i)}$$

- usual practice : posit a model for $\mathbb{P}(\xi_i = 1 | \mathbf{W}_i)$ and estimate it from all n individuals

  $\longrightarrow$ misspecification issue

1 Poisson regression model and the problem

2 The multiple imputation estimator

3 A simulation study

4 Concluding remarks

Simulation design

- simulate n individuals with $Y_i \sim \mathcal{P}(\lambda_i)$ and

$$\lambda_i = \exp(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i}),$$

with $X_{2i} \sim \mathcal{N}(0,1)$, $X_{3i} \sim \mathcal{B}(1,0.3)$, $X_{4i} \sim \mathcal{N}(0,1.5)$, $X_{5i} \sim \mathcal{U}[2,5]$ and $\beta = (0.2, -0.1, 0.4, 0.3, 0.5)$

- censoring mechanism :

$$\text{logit}(m(\mathbf{W}_i, \theta)) = \theta_1 + \theta_2 X_{2i} + \theta_3 X_{3i} + \theta_4 X_{4i} + \theta_5 X_{5i} + \theta_6 Y_i$$

- missingness mechanism :

$$\text{logit}(\mathbb{P}(\xi_i = 1 | \mathbf{W}_i)) = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{4i} + \gamma_5 X_{5i} + \gamma_6 Y_i^*$$

- various n, censoring and missingness proportions
- we take $M = 50$

Results (based on 1000 simulated samples)

FD : full data, CC : complete-case ; censoring : 20%, missing δ : 40%

		correct $m(\mathbf{W},\theta)$					incorrect $m(\mathbf{W},\theta)$				
		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
FD	bias	-0.0015	-0.0001	-0.0001	0.0006	0.0003	-0.0015	-0.0001	-0.0001	0.0006	0.0003
	SE	0.0765	0.0150	0.0323	0.0115	0.0188	0.0765	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1100	0.0213	0.0447	0.0162	0.0270	0.1100	0.0213	0.0447	0.0162	0.0270
	CP	0.9370	0.9450	0.9520	0.9540	0.9440	0.9370	0.9450	0.9520	0.9540	0.9440
CC	bias	0.1224	0.0121	-0.0128	-0.0177	-0.0152	0.1224	0.0121	-0.0128	-0.0177	-0.0152
	SE	0.0993	0.0186	0.0391	0.0151	0.0233	0.0993	0.0186	0.0391	0.0151	0.0233
	RMSE	0.1864	0.0289	0.0562	0.0275	0.0363	0.1864	0.0289	0.0562	0.0275	0.0363
	CP	0.7500	0.8940	0.9400	0.7800	0.8920	0.7500	0.8940	0.9400	0.7800	0.8920
MI	bias	0.0056	0.0004	-0.0018	-0.0005	-0.0011	0.0763	0.0010	-0.0168	-0.0110	-0.0177
	SE	0.0777	0.0150	0.0326	0.0116	0.0191	0.0822	0.0156	0.0352	0.0125	0.0200
	RMSE	0.1124	0.0215	0.0458	0.0165	0.0278	0.1412	0.0222	0.0514	0.0209	0.0340
	CP	0.9440	0.9420	0.9440	0.9470	0.9420	0.8330	0.9450	0.9390	0.8690	0.8440

Incorrect $m(\mathbf{W},\theta)$: $\text{logit}(m(\mathbf{W}_i,\theta)) = \theta_1 + \theta_2 X_{2i} + \theta_3 X_{3i} + \theta_4 Y_i^*$

1 Poisson regression model and the problem

2 The multiple imputation estimator

3 A simulation study

4 Concluding remarks

Concluding remarks

- nonparametric or semiparametric estimation of the model for missing data
- zero-inflated data
- interval-censoring

Thanks for your attention !

Concluding remarks

- nonparametric or semiparametric estimation of the model for missing data
- zero-inflated data
- interval-censoring

Thanks for your attention !