Zero-inflated Poisson regression with right-censored data

V. T. Nguyen, J.-F. Dupuy

Institut de Recherche Mathématique, Univ. Rennes

Journées du GDR Statistique et Santé
Nantes, 28/09/2018
National Medical Expenditure Survey (NMES)

US survey on medical spending (1987-88):

- \( n = 4406 \) individuals, aged \( \geq 66 \), covered by Medicare
- several recorded counts:
  - office visits and outpatient appointments to physicians or non-physician health professionals
  - emergency care, ...
- and explanatory variables:
  - demographic variables: gender, age
  - socio-economic variables: educational level, family income, insurance
  - health status measures: number of chronic conditions, self-perceived health level

Objective: model healthcare consumption, identify determinants of healthcare renunciation
Figure 1 – Frequency distributions of the number of various types of appointments.
Zero-inflation

⇒ a large number of observations of the value 0, whatever type of healthcare ⇔ zero-inflation (to be tested)

Common phenomenon in many fields, e.g.:

- **car insurance**: number of at-fault accidents declared in an insurance portfolio, due to no-claims bonus,
- **healthcare consumption**: numbers of visits to a physician, of medical prescriptions, of medical leaves...over a given period of time

Zero-inflation is a cause for overdispersion.
Zeros are assumed to arise in two ways corresponding to distinct underlying states:

- the first state occurs with probability $\omega$ and produces only zeros (structural zeros),

  E.g., individuals who have systematically decided never to visit a physician

- the other state occurs with probability $(1 - \omega)$ and is driven by a standard Poisson distribution (random zeros).

  E.g., individuals who are prepared to visit a physician but never needed to over the study period
This two-state process yields a two-component mixture distribution:

\[
Z \sim \begin{cases} 
0 & \text{with probability } \omega, \quad 0 \leq \omega \leq 1, \\
\mathcal{P}(\lambda) & \text{with probability } 1 - \omega,
\end{cases}
\]

with probability mass function

\[
P(Z = z) = \begin{cases} 
\omega + (1 - \omega)e^{-\lambda}, & z = 0 \\
(1 - \omega)e^{-\lambda}\frac{\lambda^z}{z!}, & z = 1, 2, \ldots
\end{cases}
\]

We note \(Z \sim \text{ZIP}(\lambda, \omega)\).
Zero-inflated Poisson model

Note that:

\[ P(Z = 0) = e^{-\lambda} + \omega(1 - e^{-\lambda}) \geq e^{-\lambda} = P(\mathcal{P}(\lambda) = 0) \]

Note also that

\[ E(Z) = (1 - \omega)\lambda, \]

and

\[ \text{var}(Z) = (1 + \omega \lambda)E(Z) > E(Z) \]

whenever \( \omega > 0 \) ⇒ zero inflation is a cause of overdispersion.
Zero-inflated Poisson model

Various test statistics for Poisson vs ZIP (i.e. for $H_0 : \omega = 0$),
e.g.:

- **score tests**: van den Broek, 1995; Jansakul and Hinde, 2002,
- **Wald, LR tests**: Jansakul and Hinde, 2002; Min and Czado, 2010.

**Remarks**:

- In R: van den Broek’s score test in zitest (countreg). LR test easily coded.
- Under the alternative of a ZIP model (i.e. $H_1 : \omega > 0$), $H_0$ corresponds to $\omega$ being on the parameter space boundary.
  ⇒ null asymptotic distribution is an equal mixture of $\delta_0$ and a $\chi^2$
- All tests significant in the examples above.
ZIP regression model

Lambert (1992) suggests the following models for $\lambda$ and $\omega$:

$$\log(\lambda) := \beta^\top X$$

and

$$\text{logit}(\omega) := \log \left( \frac{\omega}{1 - \omega} \right) = \gamma^\top W$$

with

- $X = (1, X_2, \ldots, X_p)^\top$ and $W = (1, W_2, \ldots, W_q)^\top$ some observed covariables,
- $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ unknown regression parameters.
→ independent observations of

\[ Z_i \sim \begin{cases} 
0 & \text{with probability } \omega_i, \\
\mathcal{P}(\lambda_i) & \text{with probability } 1 - \omega_i, i = 1, \ldots, n,
\end{cases} \]

with \( \logit(\omega_i) = \gamma^\top W_i \) and \( \log(\lambda_i) = \beta^\top X_i \).

- likelihood of \( \psi := (\beta, \gamma) : \)

\[
L_n(\psi) = \prod_{i=1}^{n} \left( \omega_i + (1 - \omega_i)e^{-\lambda_i} \right)^{\{Z_i=0\}} \cdot \left( (1 - \omega_i)e^{-\lambda_i}\frac{\lambda_i^{Z_i}}{Z_i!} \right)^{\{Z_i>0\}}
\]

- maximum likelihood estimator is consistent and asymptotically normally distributed (Erhardt, 2006; Czado et al., 2007).
The count $Z_i$ is right-censored if the true count is higher than the observed one.

- E.g.: the number of visits to a physician is right-censored at $C$ if we only know that the true number is greater than $C$.

**Modelling:**

- censoring random variable $C_i$
- define $Z_i^* = \min(Z_i, C_i)$ and $\delta_i = 1\{Z_i < C_i\}$

  (if $Z_i = C_i$, we let $Z_i^* = C_i$ and $\delta_i = 0$)

**Observations:** $n$ independent vectors $(Z_i^*, \delta_i, X_i, W_i)$ (in the complete case, we have $(Z_i, X_i, W_i)$)
Estimation

\[ L_n(\beta, \gamma) = \prod_{i=1}^{n} \begin{pmatrix} \mathbb{P}(Z_i = Z_i^* | X_i, W_i) \delta_i \mathbb{P}(Z_i \geq Z_i^* | X_i, W_i)^{1-\delta_i} \\
\end{pmatrix} \]

\[ = \prod_{i=1}^{n} \begin{pmatrix} \left( e^{-\lambda_i} \frac{Z_i^*}{Z_i^*!} (1 - \omega_i) \right)^{1-J_i} \\
(\omega_i + (1 - \omega_i) e^{-\lambda_i})^{J_i} \\
\end{pmatrix} \delta_i \\
\times \left( 1 - \sum_{k=0}^{Z_i^*-1} e^{-\lambda_i} \frac{\lambda_i^k}{k!} (1 - \omega_i) - \omega_i \right)^{(1-\delta_i)(1-J_i)} \]

with \( J_i = 1\{Z_i^* = 0\} \).

The MLE is **consistent and asymptotically normal** (asymptotic variance of \( \hat{\gamma}_n \) is the same as in the uncensored case).
Simulation study

Simulate $N = 1000$ samples from the model:

$$
\begin{align*}
\log(\lambda_i) &= \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6}, \\
\logit(\omega_i) &= \gamma_1 + \gamma_2 W_{i2} + \gamma_3 W_{i3} + \gamma_4 W_{i4} + \gamma_5 W_{i5},
\end{align*}
$$

where

- $X_{i2} \sim \mathcal{N}(0, 1)$, $X_{i3} \sim \mathcal{B}(0.3)$, $X_{i4} \sim \mathcal{N}(1, 2.25)$, $X_{i5} \sim \mathcal{E}(1)$, $X_{i6} \sim \mathcal{U}(2, 5)$, $W_{i4} \sim \mathcal{N}(-1, 1)$, $W_{i5} \sim \mathcal{B}(0.5)$
- linear predictors share common terms: $W_{i2} = X_{i2}$, $W_{i3} = X_{i3}$
- two sets of values for $\gamma$ ⇒ average fractions of ZI in the $N$ data sets: 20% and 40%
- sample size: $n = 500, 1000, 2500$
Simulation study

- \( C_i \sim \text{truncated Poisson}(\mu) \), with \( \mu \) chosen to yield average censoring proportions of 0.1, 0.2, 0.4
- Newton-Raphson algorithm (R package maxLik), with starting values obtained from a ZIP model without caring of censoring (zeroinfl in R package pscl)

Observations:

- Accuracy of MLEs of both \( \beta_j \) and \( \gamma_k \) decreases as sample size decreases
- Accuracy of MLE of \( \beta_j \) decreases as censoring increases. MLE of \( \gamma \) not sensitive to censoring
- For given \( c \) and \( n \), MLEs of the \( \beta_j \) (resp. \( \gamma_k \)) perform better when ZI decreases (resp. increases)
Simulation results ($n = 500$, $ZI = 40\%$, $c = 0.4$)

Figure 2 – Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$. 
Simulation results \( (n = 500, ZL= 40\%, \ c = 0.4) \)

Figure 3 – Histogram of the normalized estimates \( (\hat{\beta}_{j,n} - \beta_j) / \text{s.e.}(\hat{\beta}_{j,n}) \).