Zero-inflated Poisson regression with right-censored data

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

Institut de Recherche Mathématique de Rennes & INSA

Conference on "Multivariate Count Analysis"
LMB, Besançon, July 4-6 2018

Outline

1. Some background on ZIP model
2. ZIP regression with censored data
3. Simulation study
4. National Medical Expenditure Survey data
5. Concluding remarks

A real data example

Data from the US National Medical Expenditure Survey (NMES)

- national survey on medical spending (1987-88)
- $n = 4406$ individuals, aged 66 and over, covered by Medicare [1]
- for each of them, several counts are recorded :
  - number of office visits to a physician, to a non-physician health professional (nurse, physical or occupational therapists. . .)
  - number of outpatient appointments with a physician, a non-physician
  - emergency care, . . .
- **Objective** : model healthcare consumption

Now available as `NMES1988` in the `R` package `AER`. Originally taken from Deb and Trivedi (J. Applied Econometrics, 1997).

1. a health insurance program managed by the US federal government on behalf of over-65s

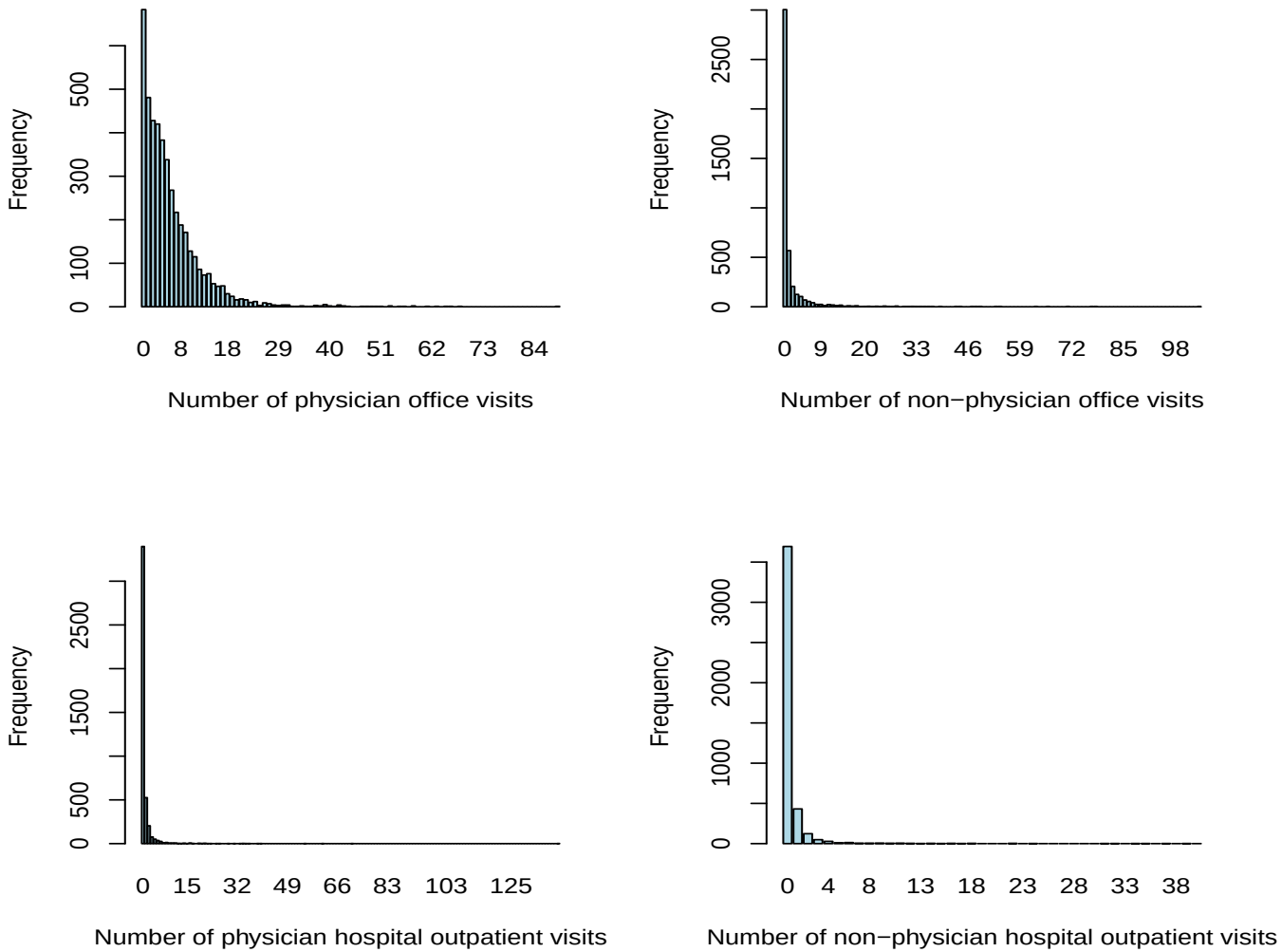

FIGURE 1 – Frequency distributions of the number of various types of appointments.

A real data example

Available explanatory variables :

- demographic variables : `gender, age`,
- socio-economic variables : `marital status, educational level, family income`, indicators of whether individual is covered by Medicaid [2] and a supplemental private insurance,
- health status measures : `number of chronic conditions` (arthritis, diabetes...) and `self-perceived health level` (poor, average, excellent).

Objectives : relate covariables to healthcare consumption and identify determinants of healthcare renunciation

2. a US health insurance for individuals with low resources

Zero-inflation, overdispersion

Figure 1 reveals a large number of observations of the value 0, whatever type of healthcare $\hookrightarrow$ zero-inflation (to be tested next)

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.

Zero-inflation, overdispersion

Poisson distribution is equidispersed. If $Z \sim \mathcal{P}(\lambda(\mathbf{X}))$, with $\mathbf{X}$ a set of covariables, equidispersion means $\mathbb{E}(Z|\mathbf{X}) = \text{var}(Z|\mathbf{X})$.

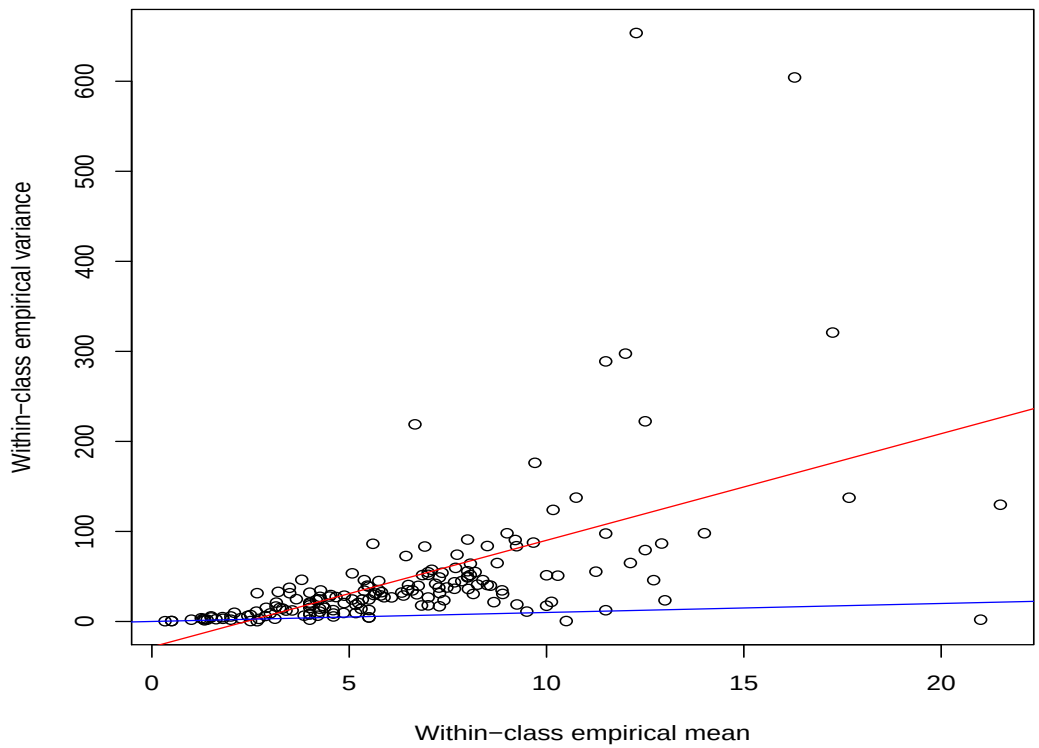

FIGURE 2 – Empirical mean and variance of the number of physician office visits, after segmentation of `health`, `med`, age, and `numchron`.

Zero-inflation, overdispersion

Overdispersion can arise for various reasons :

- presence of unobserved heterogeneity in the data,
- omission of one or several key covariables in the model,
- zero inflation,
- ...

Zero-inflation

Zeros are assumed to arise in two ways corresponding to distinct underlying states :

- the first state occurs with probability ω 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

Zero-inflation

Zeros are assumed to arise in two ways corresponding to distinct underlying states :

- the first state occurs with probability ω 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

Zero-inflation

Zeros are assumed to arise in two ways corresponding to distinct underlying states :

- the first state occurs with probability ω 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

Zero-inflated Poisson model

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

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

We note $Z \sim \mathsf{ZIP}(\lambda, \omega)$.

Zero-inflated Poisson model

Note that :

$$\begin{aligned}\mathbb{P}(Z=0) &= e^{-\lambda}+\omega(1-e^{-\lambda}) \\ &\geq e^{-\lambda}=\mathbb{P}(\mathcal{P}(\lambda)=0)\end{aligned}$$

Note also that

$$\mathbb{E}(Z)=(1-\omega)\lambda,$$

and

$$\text{var}(Z)=(1+\omega\lambda)\mathbb{E}(Z)>\mathbb{E}(Z)$$

whenever $\omega > 0$. Check that zero inflation is a cause of overdispersion.

Zero-inflated Poisson model

Various test statistics for Poisson vs ZIP (i.e. for $H_0 : \omega = 0$) have been proposed, such as

- score tests : van den Broek (1995) for constant ω ; Jansakul and Hinde (2002) for covariate-dependent ω,
- Wald and likelihood ratio 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 ω being on the parameter space boundary.

  $\Rightarrow$ the null asymptotic distribution is an equal mixture of δ_0 and a χ^2
- All tests significant in the examples above (LR based on ZIP regression model).

Zero-inflated Poisson model

Various test statistics for Poisson vs ZIP (i.e. for $H_0 : \omega = 0$) have been proposed, such as

- score tests : van den Broek (1995) for constant ω ; Jansakul and Hinde (2002) for covariate-dependent ω,
- Wald and likelihood ratio 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 ω being on the parameter space boundary.

  $\Rightarrow$ the null asymptotic distribution is an equal mixture of δ_0 and a χ^2
- All tests significant in the examples above (LR based on ZIP regression model).

Zero-inflated Poisson model

Various test statistics for Poisson vs ZIP (i.e. for $H_0 : \omega = 0$) have been proposed, such as

- score tests : van den Broek (1995) for constant ω ; Jansakul and Hinde (2002) for covariate-dependent ω,
- Wald and likelihood ratio 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 ω being on the parameter space boundary.

  $\Rightarrow$ the null asymptotic distribution is an equal mixture of δ_0 and a χ^2
- All tests significant in the examples above (LR based on ZIP regression model).

ZIP regression model

Lambert (1992) suggests the following models for λ and ω :

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

and

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

with

- $\mathbf{X} = (1, X_2, \ldots, X_p)^\top$ and $\mathbf{W} = (1, W_2, \ldots, W_q)^\top$ vectors of observable covariables,
- $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ unknown regression parameters

Since then, the model has been extended in several directions, e.g. ZIP model with cluster specific random effects, semiparametric and doubly semiparametric ZIP models. . .

Estimation and asymptotics

Consider n independent observations of the ZIP regression model

$$Z_i \sim \omega_i \delta_0 + (1-\omega_i)\mathcal{P}(\lambda_i), i = 1, \ldots, n,$$

with $\text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i$ and $\log(\lambda_i) = \beta^\top \mathbf{X}_i$.

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

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

- the maximum likelihood estimator

$$\hat{\psi}_n = \arg\max_{\psi} L_n(\psi)$$

is consistent and asymptotically normally distributed under some regularity conditions (Erhardt, 2006 ; Czado et al., 2007).

Right-censored observations

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 $(Z_i^*, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$ (in the complete case, we have $(Z_i, \mathbf{X}_i, \mathbf{W}_i)$)

Estimation

Likelihood of $\psi := (\beta^\top, \gamma^\top)^\top$:

$$
\begin{aligned}
L_n(\psi) &= \prod_{i=1}^{n} \mathbb{P}(Z_i = Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{\delta_i} \mathbb{P}(Z_i \geq Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\left(e^{-\lambda_i} \frac{\lambda_i^{Z_i^*}}{Z_i^*!} (1-\omega_i) \right)^{1-J_i} \left(\omega_i + (1-\omega_i) e^{-\lambda_i} \right)^{J_i} \right)^{\delta_i} \\
&\qquad \times \left(1 - \sum_{k=0}^{Z_i^*-1} e^{-\lambda_i} \frac{\lambda_i^k}{k!} (1-\omega_i) - \omega_i \right)^{(1-\delta_i)(1-J_i)}
\end{aligned}
$$

with $J_i = 1_{\{Z_i^*=0\}}$, $\text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i$ and $\log(\lambda_i) = \beta^\top \mathbf{X}_i$.

MLE : $\hat{\psi}_n = \arg\max_\psi \ell_n(\psi)$ with $\ell_n = \log L_n$.

Model assumptions

Assume that

C1 Covariates are bounded, i.e., there exist compact sets $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{W} \subset \mathbb{R}^q$ such that $\mathbf{X}_i \in \mathcal{X}$ and $\mathbf{W}_i \in \mathcal{W}$ $\forall i = 1, 2, \ldots$

C2 The true parameter value $\psi_0 = (\beta_0^\top, \gamma_0^\top)^\top$ lies in the interior of some known compact set $\mathcal{C} = \mathcal{B} \times \mathcal{G} \subset \mathbb{R}^k$ (where $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{G} \subset \mathbb{R}^q$ are the parameter spaces of β and γ respectively).

C3 There exists a constant $c_1 > 0$ such that $n/\lambda_{\min}(F_n(\psi_0)) \leq c_1$ for every $n = 1, 2, \ldots$ with $F_n(\psi) = -\mathbb{E}(\partial^2 \ell_n(\psi)/\partial\psi\partial\psi^\top)$.

C4 Censoring random variables $C_i, i = 1, 2, \ldots$ are strictly positive and bounded by some constant $M < \infty$.

Asymptotics

Theorem 1

Under conditions C1-C4, as $n \to \infty$:

- *$\hat{\psi}_n$ converges in probability to ψ_0,*
- *let $S_n(\psi) = \partial \ell_n(\psi)/\partial \psi$, then $F_n^{-\frac{1}{2}}(\hat{\psi}_n) S_n$ converges in distribution to $\mathcal{N}(0, I_k)$,*
- *$F_n^{1/2}(\hat{\psi}_n)(\hat{\psi}_n - \psi_0)$ converges in distribution to $\mathcal{N}(0, I_k)$.*

Remark : asymptotic variance of $\hat{\gamma}_n$ is the same as in the uncensored case.

$\Rightarrow$ only accuracy of $\hat{\beta}_n$ might be affected by censoring (note however that $\hat{\beta}_n$ is involved in the asymptotic variance of $\hat{\gamma}_n$).

Simulation design

Consider the model

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

where

- $X_{i2} \sim \mathcal{N}(0,1)$, $X_{i3} \sim \mathcal{B}(0.3)$, $X_{i4} \sim \mathcal{N}(1, 2.25)$, $X_{i5} \sim \mathcal{E}(1)$, $X_{i6} \sim \mathcal{U}(2,5)$, $W_{i4} \sim \mathcal{N}(-1,1)$, $W_{i5} \sim \mathcal{B}(0.5)$ (all indep.)
- linear predictors share common terms, with $W_{i2} = X_{i2}$ and $W_{i3} = X_{i3}$
- $\beta = (0.7, 0.1, 0.4, 0.85, -0.5, 0)$
- sample size : $n = 500, 1000, 2500$

Simulation design

- simulate $N = 1000$ samples
- two cases :
  - $\gamma = (-0.9, -0.65, -0.2, 0.65, 0)$ (average fraction of ZI in the N data sets is 20%)
  - $\gamma = (0.25, -0.7, -0.2, 0.65, 0)$ (average fraction of ZI is 40%)
- $C_i \sim$ truncated Poisson(μ), with μ chosen to yield average censoring proportions $c = 0.1, 0.2, 0.4$ in the N samples
- Newton-Raphson algorithm (`R` package `maxLik`), with starting values obtained by estimating a ZIP model without caring of censoring (`zeroinfl` in `R` package `pscl`)

Simulation results

Several accuracy measures :

- average relative bias of the estimates $\hat{\beta}_{j,n}$ and $\hat{\gamma}_{k,n}$ over the N simulated samples, e.g. :

$$\frac{1}{N}\sum_{t=1}^{N}\frac{\hat{\beta}_{j,n}^{(t)}-\beta_j}{\beta_j}\times 100,$$

  with $\hat{\beta}_{j,n}^{(t)}$ the MLE of β_j in the t-th simulated sample, $t=1,\ldots,N$

- average standard error and root mean square error
- average length of 95%-level CI and their empirical coverage probability

Simulation results

Various plots :

- normal Q-Q plots of the estimates
- histograms of the normalized estimates, e.g. :

$$\frac{\widehat{\beta}_{j,n} - \beta_j}{\text{standard error}(\widehat{\beta}_{j,n})}$$

and

$$\frac{\widehat{\gamma}_{j,n} - \gamma_j}{\text{standard error}(\widehat{\gamma}_{j,n})}$$

Simulation results ($n = 500$, ZI$=$ 40%)

		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
c		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	rel. bias	0.3032	-0.5554	0.3613	**0.0162**	-0.1714	-	2.0303	**2.3779**	4.4416	1.8185	-
	SE	0.1038	0.0226	0.0437	**0.0158**	0.0323	0.0247	0.2155	**0.1315**	0.2546	0.1294	0.2349
	RMSE	0.1473	0.0323	0.0621	**0.0222**	0.0465	0.0349	0.3103	**0.1897**	0.3593	0.1841	0.3388
	CP	0.9530	0.9500	0.9480	**0.9530**	0.9390	0.9520	0.9460	**0.9530**	0.9540	0.9460	0.9450
	ℓ	0.4056	0.0883	0.1708	**0.0614**	0.1260	0.0967	0.8438	**0.5144**	0.9973	0.5061	0.9205
0.1	rel. bias	-0.5502	-0.1754	0.3451	**0.5726**	0.2389	-	1.1946	**2.4539**	4.6880	1.9284	-
	SE	0.1506	0.0341	0.0689	**0.0357**	0.0438	0.0370	0.2164	**0.1320**	0.2556	0.1296	0.2352
	RMSE	0.2123	0.0482	0.0981	**0.0501**	0.0634	0.0523	0.3112	**0.1904**	0.3604	0.1845	0.3391
	CP	0.9440	0.9490	0.9430	**0.9560**	0.9440	0.9440	0.9470	**0.9540**	0.9520	0.9470	0.9470
	ℓ	0.5894	0.1334	0.2698	**0.1399**	0.1709	0.1450	0.8475	**0.5165**	1.0012	0.5069	0.9216
0.2	rel. bias	-0.9652	1.6303	1.7772	**1.1197**	0.8294	-	0.5822	**2.4105**	3.9740	1.9938	-
	SE	0.1906	0.0448	0.0915	**0.0489**	0.0542	0.0483	0.2170	**0.1326**	0.2565	0.1297	0.2354
	RMSE	0.2702	0.0636	0.1276	**0.0695**	0.0777	0.0683	0.3118	**0.1919**	0.3616	0.1847	0.3392
	CP	0.9470	0.9490	0.9600	**0.9530**	0.9460	0.9480	0.9430	**0.9550**	0.9550	0.9460	0.9480
	ℓ	0.7457	0.1750	0.3579	**0.1912**	0.2117	0.1889	0.8496	**0.5186**	1.0047	0.5074	0.9222
0.4	rel. bias	-1.5772	7.8277	6.3481	**3.5830**	3.0722	-	1.1201	**2.2840**	2.9409	2.2563	-
	SE	0.3550	0.0887	0.1850	**0.0912**	0.0924	0.0932	0.2191	**0.1361**	0.2627	0.1302	0.2359
	RMSE	0.5040	0.1293	0.2649	**0.1335**	0.1330	0.1327	0.3149	**0.1964**	0.3719	0.1857	0.3402
	CP	0.9520	0.9410	0.9490	**0.9430**	0.9480	0.9450	0.9490	**0.9490**	0.9530	0.9510	0.9470
	ℓ	1.3864	0.3456	0.7212	**0.3555**	0.3601	0.3643	0.8578	**0.5321**	1.0287	0.5092	0.9241

Simulation results

Making n and ZI vary, we observe that :

- accuracy of MLEs of both β_j and γ_k decreases as sample size decreases,
- for given c and n, MLEs of the β_j perform better when ZI decreases,
- for given c and n, MLEs of the γ_k perform better when ZI increases.

Simulation results ($n = 500$, ZI$= 40\%$, $c = 0.4$)

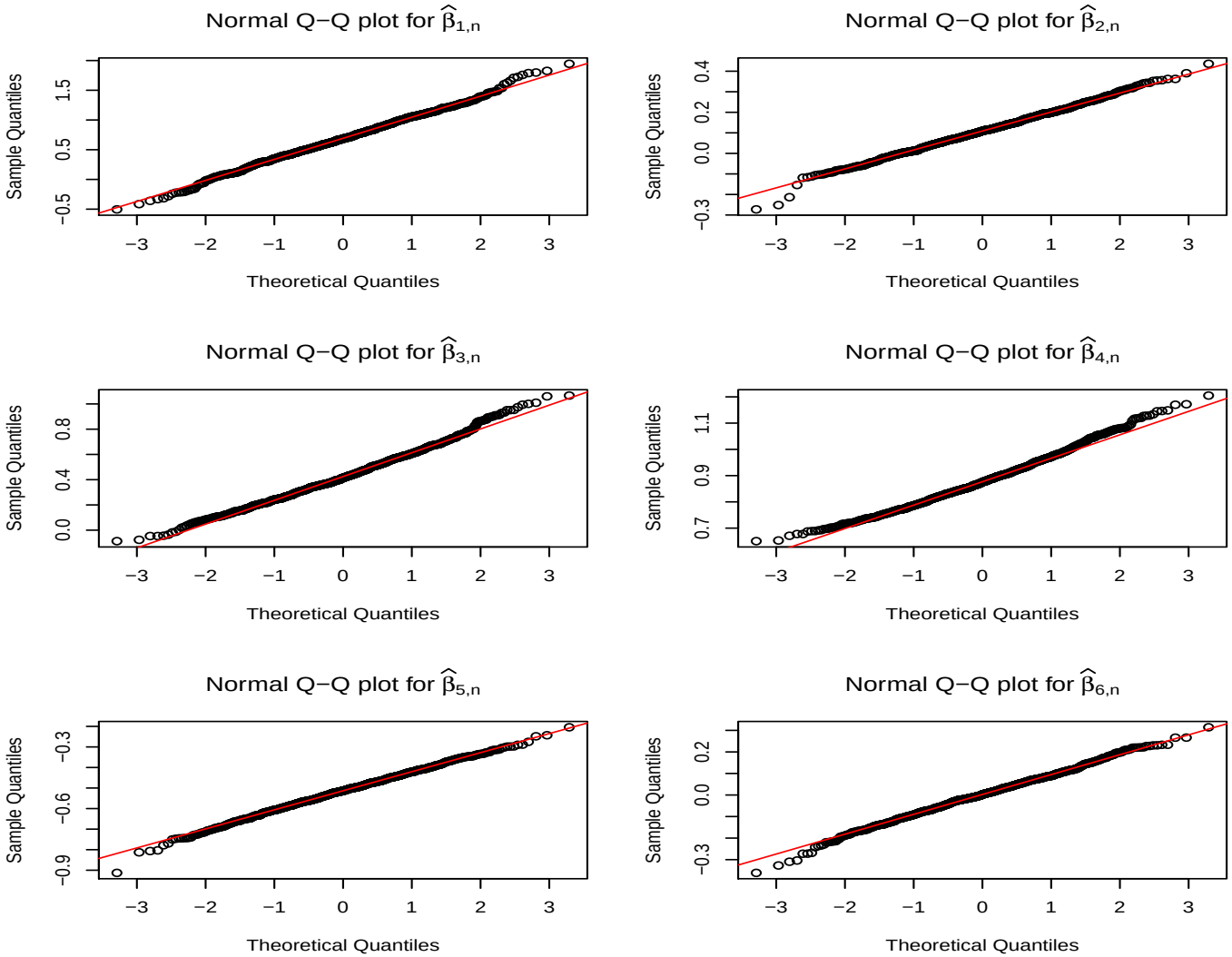

FIGURE 3 – Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$.

Simulation results ($n = 500$, ZI$= 40\%$, $c = 0.4$)

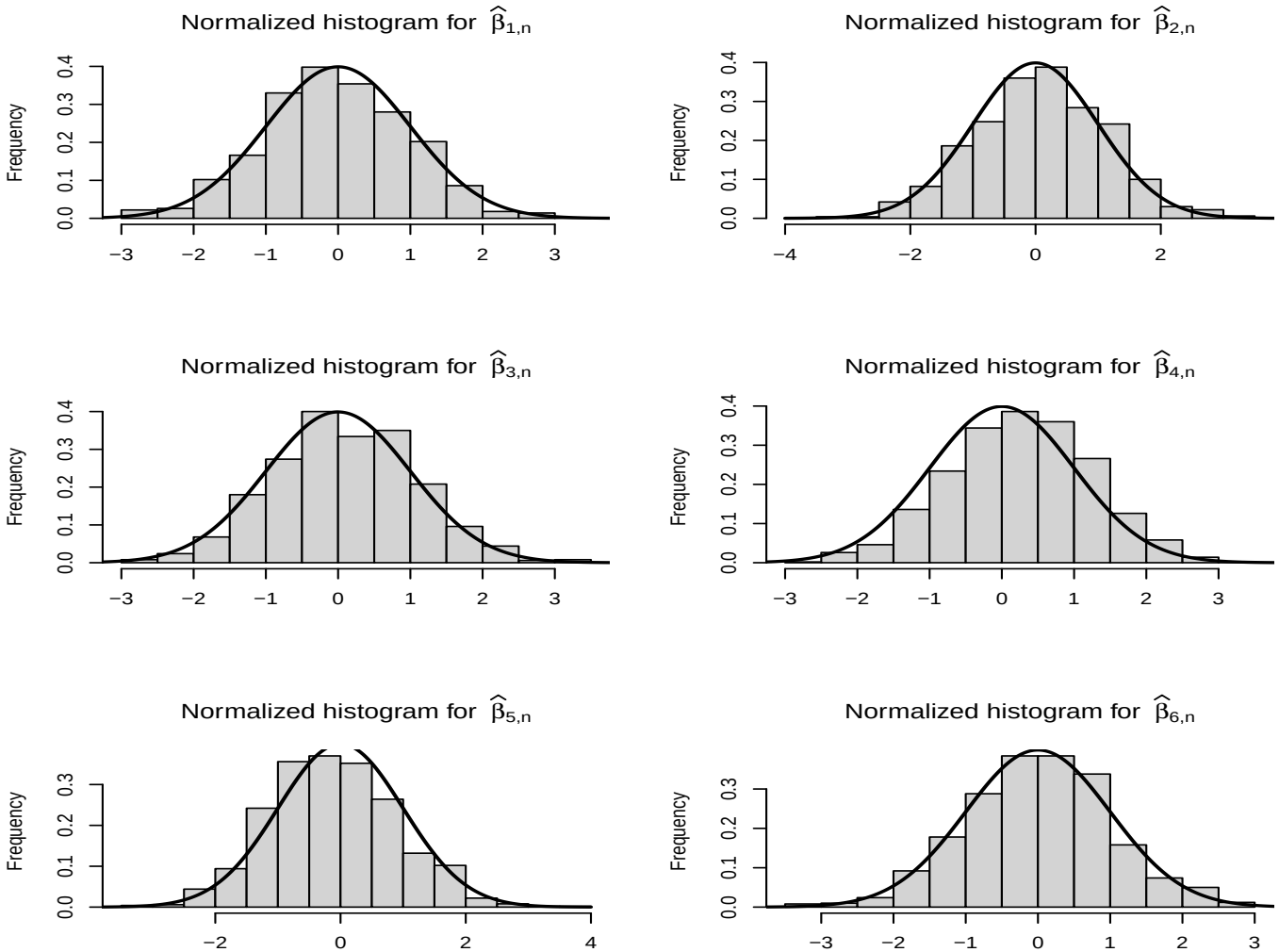

FIGURE 4 – Histogram of the normalized estimates $(\widehat{\beta}_{j,n} - \beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$.

NMES analysis

National survey on medical spending (1987-88) conducted on $n = 4406$ individuals, aged 66 and over, covered by Medicare.

We consider

- the number of office visits to a physician as the response variable (uncensored in the initial dataset, here censored at level 0.2 and 0.4)
- the following explanatory variables :
  - gender (1 for female, 0 for male)
  - age
  - marital status (1 if married, 0 otherwise)
  - educational level (number of years of education)
  - family income
  - indicators of Medicaid/supplemental private insurance (1 if covered, 0 otherwise)
  - number of chronic conditions
  - self-perceived health level (poor, average, excellent)

NMES analysis

We recode self-perceived health level (poor, average, excellent) as

- `health1` (1 if perceived as poor, 0 otherwise),
- `health2` (1 if perceived as excellent, 0 otherwise).

The fitted model is :

$$\begin{cases} \log(\lambda_i) = \beta_1 + \beta_2\,\texttt{gender}_i + \beta_3\,\texttt{age}_i + \beta_4\,\texttt{marital\_status}_i + \beta_5\,\texttt{school}_i \\ \qquad\qquad +\beta_6\,\texttt{income}_i + \beta_7\,\texttt{medicaid}_i + \beta_8\,\texttt{insurance}_i + \beta_9\,\texttt{chronic}_i \\ \qquad\qquad\quad +\beta_{10}\,\texttt{health1}_i + \beta_{11}\,\texttt{health2}_i, \\ \\ \text{logit}(\omega_i) = \gamma_1 + \gamma_2\,\texttt{gender}_i + \gamma_3\,\texttt{age}_i + \gamma_4\,\texttt{marital\_status}_i + \gamma_5\,\texttt{school}_i \\ \qquad\qquad +\gamma_6\,\texttt{income}_i + \gamma_7\,\texttt{medicaid}_i + \gamma_8\,\texttt{insurance}_i + \gamma_9\,\texttt{chronic}_i \\ \qquad\qquad\quad +\gamma_{10}\,\texttt{health1}_i + \gamma_{11}\,\texttt{health2}_i. \end{cases}$$

	no censoring		censoring = 20%		censoring = 40%	
variable	estimate (s.e.)	signif.	estimate (s.e.)	signif.	estimate (s.e.)	signif.
			Poisson model coefficients			
intercept	1.7596 (0.0878)	***	1.3491 (0.2057)	***	1.1464 (0.1246)	***
gender	0.0066 (0.0144)		0.0487 (0.0165)	**	0.0377 (0.0201)	.
age	-0.0593 (0.0107)	***	-0.0148 (0.0248)		-0.0102 (0.0152)	
marital status	-0.0796 (0.0148)	***	-0.0269 (0.0140)	.	-0.0072 (0.0209)	
school	0.0209 (0.0020)	***	0.0130 (0.0015)	***	0.0144 (0.0027)	***
income	-0.0014 (0.0023)		-0.0008 (0.0026)		-0.0038 (0.0032)	
medicaid	0.2256 (0.0254)	***	0.1799 (0.0297)	***	0.1681 (0.0369)	***
insurance	0.1892 (0.0200)	***	0.1387 (0.0228)	***	0.1638 (0.0274)	***
chronic	0.1198 (0.0046)	***	0.1150 (0.0056)	***	0.1191 (0.0070)	***
health1	0.3081 (0.0175)	***	0.2285 (0.0212)	***	0.1971 (0.0272)	***
health2	-0.3224 (0.0312)	***	-0.2532 (0.0338)	***	-0.2018 (0.0383)	***
			Zero-inflation model coefficients			
intercept	1.9786 (0.5917)	***	1.9597 (0.6870)	**	1.9709 (0.7531)	**
gender	-0.4721 (0.0974)	***	-0.4716 (0.0997)	***	-0.4840 (0.1020)	***
age	-0.1829 (0.0741)	*	-0.1818 (0.0862)	*	-0.1876 (0.0947)	*
marital status	-0.2843 (0.1033)	**	-0.2862 (0.1060)	**	-0.2921 (0.1087)	**
school	-0.0607 (0.0128)	***	-0.0614 (0.0131)	***	-0.0611 (0.0136)	***
income	-0.0104 (0.0188)		-0.0104 (0.0193)		-0.0120 (0.0201)	
medicaid	-0.4921 (0.1713)	**	-0.4928 (0.1718)	**	-0.4875 (0.1767)	**
insurance	-0.8227 (0.1102)	***	-0.8299 (0.1099)	***	-0.8272 (0.1145)	***
chronic	-0.5414 (0.0458)	***	-0.5393 (0.0464)	***	-0.5360 (0.0479)	***
health1	0.0124 (0.1615)		0.0212 (0.1605)		0.0380 (0.1669)	
health2	0.2447 (0.1505)		0.2398 (0.1594)		0.2410 (0.1575)	

TABLE 1 – Wald test code : *** signif. at 0.1% level, ** signif. at 1% level, * signif. at 5% level, . signif. at 10% level.

NMES analysis

Censored ZIP regression model retains main conclusions of the uncensored case. E.g.,

- an increase in the number of chronic diseases increases both
  - probability of visiting a doctor
  - average number of consultations
- self-perceived health does not affect decision of visiting a doctor but affects the average number of visits
- income is non-significant in both submodels for ZI and visits frequency
- Medicaid and private insurance are significant determinants of both decision of visiting a doctor and number of visits

Concluding remarks

- we have established asymptotic properties of the MLE in censored ZIP regression ($\Rightarrow$ rigorous basis for Wald tests)
- estimation and variable selection in ZI model part are only marginally affected by censoring
- estimation and variable selection in the Poisson model part can be affected by censoring
- further research : consider more general models (e.g., semi-parametric ZIP, ZI generalized Poisson regression model,. . .)

Main references


Cameron, A. C., Trivedi, P. K., 2013.
Regression Analysis of Count Data .
Cambridge University Press, Cambridge.


Czado, C., Erhardt, V., Min, A., Wagner, S., 2007.
Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates.
Statistical Modelling 7(2), 125-153.


Fahrmeir, L., Kaufmann, H., 1985.
Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models.
The Annals of Statistics 13(1), 342-368.


Kleiber, C., Zeileis, A., 2008.
Applied Econometrics with R.
Springer-Verlag, New York.
http ://CRAN.R-project.org/package=AER.


Lambert, D., 1992.
Zero-inflated Poisson regression, with an application to defects in manufacturing.
Technometrics 34, 1-14.

Thanks for attention !