Distributional Modelling in R

Case Studies II – Discrete Distributions

Thomas Kneib, Nikolaus Umlauf

https://nikum.org/dmr.html

Count Data Regression

- Standard approach: Log-linear Poisson models embedded in the generalized linear model framework where

$$y_i \sim \text{Po}(\lambda_i), \qquad \text{and} \qquad \lambda_i = \exp(\boldsymbol{x}_i' \gamma).$$

- In many practical applications, we observe one or both of the following challenges:
  - Zero inflation, i.e. excess of zeros compared to standard count data distributions such as Poisson.
  - Overdispersion, i.e. variances of the responses exceed the expectation (unlike in Poisson regression).

Zero-Inflated Count Data Regression

- Basic idea of zero-inflated count data regression models: Zeros may arise from either
  - structural zeros, i.e. observations that are “always zero” and,
  - zeros arising from the count data distribution.

- If y_i is a count data response, assume that y_i is generated as

$$y_i = \kappa_i \tilde{y}_i$$

  where κ_i is a binary indicator for structural zeros, i.e.

$$\kappa_i \sim \mathrm{Be}(1 - \pi_i)$$

  and $\tilde{y}_i$ follows a standard count data distribution such as Poisson or negative binomial.

Interpretation

- If the binary indicator κ_i is zero, we always obtain zero as response (structural zeros) $\Rightarrow$ Structural zeros occur with probability π_i.
- If the binary indicator κ_i is one, y_i is realized from the count data model.
- Mixed density for the responses y_i:

$$p(y_i) = \pi_i \mathbb{1}_{\{0\}}(y_i) + (1 - \pi_i)\tilde{p}(y_i)$$

  where $\tilde{p}(y_i)$ is the density for the count data distribution.
- For the zero-inflated distribution we obtain

$$E(y_i) = (1 - \pi_i)E(\tilde{y}_i)$$

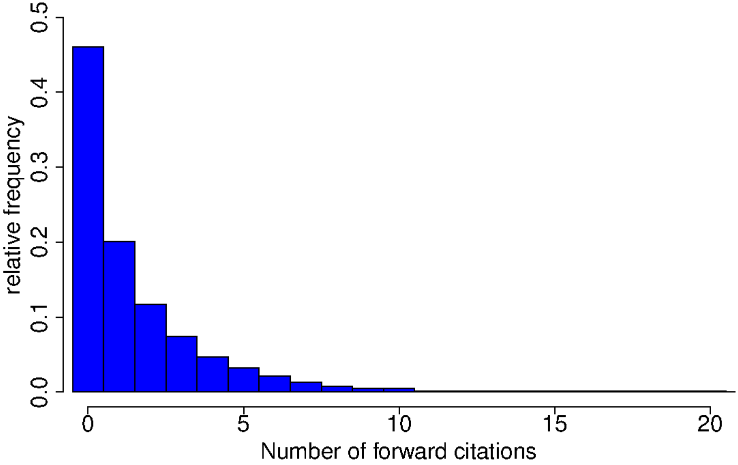
  and

$$\text{Var}(y_i) = (1 - \pi_i)\text{Var}(\tilde{y}_i) + \pi_i(1 - \pi_i)\left[E(\tilde{y}_i)\right]^2 .$$

Example: Patent Citations

- Information on 4,805 patents issued by the European Patent Office (EPO) between 1980 and 1997.
- Response variable of interest: number of forward citations.
- Explanatory variables include the grant year, the no. of designated states the patent applies to, and the no. of EPO claims.
- Characteristics of the response:
  - 46% zeros (many patents are never cited).
  - Maximum no. of citations: 40.
  - Average no. of citations: 1.63
  - Variance: 7.35

Frequency Histogram



Variables

Continuous covariates				
	description	mean	std	min/max
`year`	grant year	1991		1980/1997
`ncountry`	no. of designated states in Europe	7.77	4.12	1/17
`nclaims`	no. of EPO claims	12.33	8.13	1/50
Binary covariates				
	description	categories	rel freq	
`biopharm`	patent from biotech/pharma sector	yes=1	43.9%	
`ustwin`	U.S. twin exists	yes=1	61.3%	
`patus`	patentholder of the patent from U.S.	yes=1	33.2%	
`patgsgr`	owner from Switzerland, Ger or GB	yes=1	23.7%	
`opp`	oppositions	yes=1	41.1%	

Model Specification and Comparison

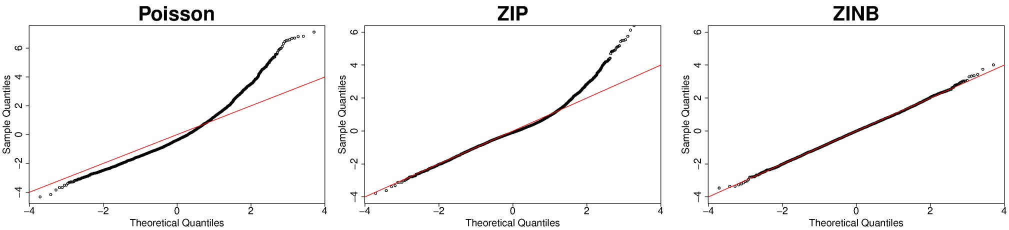
- We compare Poisson, zero-inflated Poisson and zero-inflated negative binomial models.
- Basic predictor structure for all model parameters:

$$\eta = \boldsymbol{x}'\gamma + s_1(\texttt{year}) + s_2(\texttt{ncountry}) + s_3(\texttt{nclaims}).$$

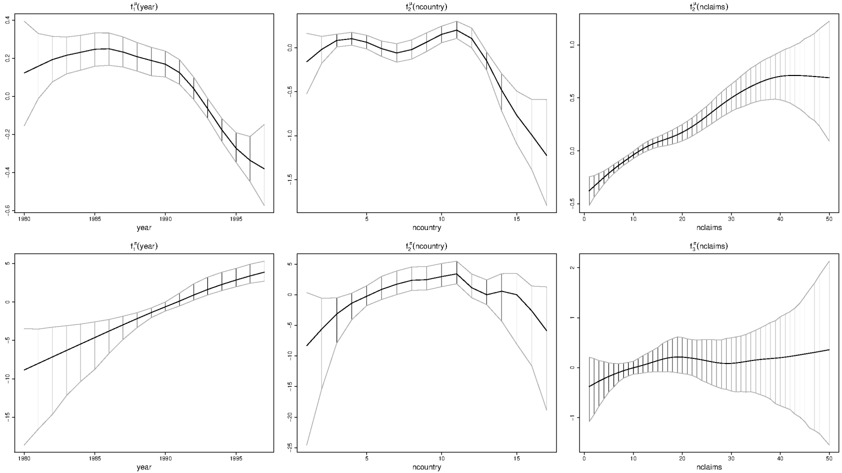
- Scores (averages from ten-fold cross-validation):

Model	Brier Score	Logarithmic Score	Spherical Score
Poisson	-0.8180	-2.3926	0.0070
ZIP	-0.7480	-2.0197	0.0077
ZINB	-0.7439	-1.7604	0.0074

Quantile Residuals



Estimated Nonlinear Effects



Estimated Nonlinear Effects