VARIABLE SELECTION STRATEGY FOR ZERO INFLATED MODELS WITH APPLICATION TO AUTOMOBILE INSURANCE DATA

Soumia KACI, Kamel BOUKHETALA, Jean-François DUPUY
USTHB Algeria, USTHB Algeria, INSA-Rennes France

ABSTRACT

When count data exhibit excess zero, that is more zero counts than a simpler parametric distribution can model, the zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB) models are often used. Variable selection for these models is even more challenging than other regression situations because the availability of p covariates implies 4^p possible models. We adapt to zero-inflated models an approach for variable selection that avoids the screening of all possible models.
As an additional novelty, we propose a new way of extracting information from a rich chain of covariates ,this approach is based on a stochastic search through all regression models with all available covariates. We fit a binary indicator of the inflation, which generates a first subset of covariates for the zero part. Poisson and Negative Binomial models are fitted to generate a second chain of covariates for the count part. Finally a backward elimination algorithm is used to fit a zero inflated model. an application on automobile insurance data is described. Finally, A simulation study is conducted to assess finite-sample behaviour, where we also compare our approach with regularization (penalized) techniques available in the literature.
Key words : excess zeros, number of claims, ZI model, variable selection.

1. INTRODUCTION

The automobile claims experience in the insurance sector is measured by the frequency of accidents and their amounts . In this highly competitive market, the insurer search to select factors that help to explain this loss experience. In business, psychology, social, and public health related research, it is common that the outcomes are relatively infrequent behaviors and phenomena. Data with abundant zeros are especially frequent in research studies when counting the occurrence of certain behavioral events, such as number of purchases made, number of schoolabsences, number of cigarettes smoked, or number of hospitalizations. These types of data are called count data and their values are usually non-negative with a lower bound of zero.
The classical Poisson regression model for count data is often of limited use in these disciplines because empirical count data sets typically exhibit over-dispersion, under-dispersion or an excess number of zeros, when, the variance is assumed to be equal to the mean, which may be violated in real data. One way to deal with over-dispersion is a negative binomial (NB) regression. The negative binomial model belongs to the family of generalized linear models [7]. However, although negative binomial model typically can capture overdispersion rather well, it is in many applications not sufficient for modeling excess zeros. In the econometrics and statistics literature, Mullahy [6] and Lambert [4] proposed the zero-inflation models that address this modeling by a second model component capturing zero counts.Zero-inflation models[4] take a somewhat different approach : they are mixture models that combine a count component and a point mass at zero. An overview of count data models in econometrics is provided in Cameron and Trivedi [2] [3].

Variable selection is important in applications because it allows to explain data in the simplest way for better interpretability (e.g., efficient identification of risk factors). It is also a mean for cost management, if the model is next used for prediction. From the statistical point of view, parsimonious models have to be preferred for better prediction accuracy and better interpretation. In addition, collinearity-related problems are mitigated when there are fewer variables involved. We focus on the more challenging ZI model even though our methodology can be applied to the hurdle model as well : for p available covariates, there are as many as 4^p possible different ZI models, whereas this number is 2×2^p for the hurdle model, due to the orthogonality of its two parts, which can be fitted separately.

The manuscript has the following structure. In Section 2, we present generalized linear models for count data. A real data analysis and our proposal for stochastic variable selection are presented in Section 3 , followed by Section 4,which introduce simulations settings, where different versions of our approach are contrasted with existing alternatives.
Throughout this paper,Y denotes the endogeneous (dependent, explained) variable, and $(\mathbf{X}, \mathbf{W})$ denotes the exogenuous (independent, explanatory) variable(s)

2. REGRESSION MODELS

2.1. Zero-inflated Poisson model

The term "zero inflation" describes a situation in which the number of zeros observed in a sample of count data is higher than the number predicted by "classical" count models (such as Poisson or binomial models). One of the most common approaches to working with this type of data is to assume that the probability distribution of the count variable (denoted as Y below) is a mixture of a degenerate distribution at zero (i.e. a distribution that takes the value 0 with probability 1) and a count model. To illustrate this idea, we suppose that the count model follows a Poisson distribution $\mathscr{P}(\lambda)$. The distribution of Y can then be written as follows :

$$Y \sim \pi\delta_0 + (1-\pi)\mathscr{P}(\lambda). \tag{1}$$

In the above expression, π is the probability that Y is systematically equal to zero (called the "zero inflation probability" below) and δ_0 denotes the degenerate distribution at zero. Equation 1 can alternatively be interpreted as follows :

$$Y \sim \begin{cases} 0 & \text{with probability} \quad \pi \\ \mathscr{P}(\lambda) & \text{with probability} \quad 1-\pi \end{cases}$$

Suppose that we are observing a count variable Y on a sample of n individuals. Let us write Y_i for the observed value of Y at the i-th individual, $i = 1, \therefore, n$. We can construct a zero-inflated Poisson regression model for Y_i by allowing the probability π and the intensity λ in equation (1) to depend on the individual i via the explanatory variables (or covariables). This model can be stated as :

$$\mathbb{P}(Y=y) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i} & z=0 \\ (1-\pi_i)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & z=1,2,.. \end{cases}$$

where π_i and λ_i are, respectively, functions of the vectors of covariables $\mathbf{W}_i = (\mathbf{W}_{i1}, ..., \mathbf{W}_{iq})^T$ and $\mathbf{W}_i = (\mathbf{X}_{i1}, ..., \mathbf{X}_{ip})^T$ (setting $\mathbf{X}_{i1} = \mathbf{W}_{i1} = 1$). The components of these vectors can be either qualitative or quantitative. The probability π_i is usually described by a logistic regression :

$$\begin{aligned} &\text{logit}(\pi_i) = \gamma^T \mathbf{W}_i = \gamma_1 + \gamma_2 \mathbf{W}_{i2} + ... + \gamma_q \mathbf{W}_{iq} \\ &\iff \pi_i = \frac{\exp(\gamma^T \mathbf{W}_i)}{1+\exp(\gamma^T \mathbf{W}_i)} \in (0,1). \end{aligned}$$

and the intensity λ_i is usually modeled by :

$$\begin{array}{l} \ln(\lambda_i) = \beta^T \mathbf{X}_i = \beta_1 + \beta_2 \mathbf{X}_{i2} + ... + \beta_p \mathbf{X}_{ip} \\ \Longleftrightarrow \lambda_i = \exp(\beta^T \mathbf{X}_i) \end{array} \tag{2}$$

where $\beta = (\beta_1, ..., \beta_p)^T$ and $\gamma = (\gamma_1, ..., \gamma_q)^T$ are vectors of unknown parameters. We can summarize this model in the form :

$$\forall i = 1, ..., n \quad \left\{ \begin{array}{l} Y_i \sim \pi_i \delta_0 + (1 - \pi_i) \mathscr{P}(\lambda_i) \\ \text{logit}(\pi_i) = \gamma^T \mathbf{W}_i \\ \ln(\lambda_i) = \beta^T \mathbf{X}_i \end{array} \right.$$

Notationally, we write that $Y_i \sim \mathbf{ZIP}(\lambda_i, \pi_i)$. Conditional on $\mathbf{X}_i$ and $\mathbf{W}_i$, the mean and variance of Y_i are, respectively, given by :

$$\mathbb{E}(Y_i | \mathbf{X}_i, \mathbf{W}_i) = (1 - \pi_i)\lambda_i = \frac{\exp(\beta^T \mathbf{X}_i)}{1 + \exp(\gamma^T \mathbf{W}_i)}$$

and

$$var(Y_i | \mathbf{X}_i, \mathbf{W}_i) = (1 + \pi_i \lambda_i)(1 - \pi_i)\lambda_i = (1 + \pi_i \lambda_i)\mathbb{E}(Y_i | \mathbf{X}_i, \mathbf{W}_i).$$

The conditional distribution of Y_i is overdispersed, since $(1 + \pi_i \lambda_i) > 1$,
and hence $var(Y_i | \mathbf{X}_i, \mathbf{W}_i) > \mathbb{E}(Y_i | \mathbf{X}_i, \mathbf{W}_i)$.

2.2. Zero-inflated Negative Binomial model

The ZINB distribution is a mixture distribution assigning a mass of π_i to "extra" zeroes and a mass of $(1 - \pi_i)$ to a negative binomial distribution, where $0 \leq \pi_i \leq 1$. Note that the negative binomial distribution is a continuous mixture of Poisson distributions, which allows the Poisson mean λ to be gamma distributed and in this way overdispersion is modelled. Observe that this distribution is also useful when the count is made of correlated binary random variables. More specifically, the negative binomial distribution is given by

$$P(Y_i = y) = \frac{\Gamma(y + \nu)}{\Gamma(\nu) y!} \left(\frac{\mu_i}{\nu + \mu_i} \right)^y \left(\frac{\nu}{\nu + \mu_i} \right)^\nu \quad y = 0, 1, 2, ..; \mu_i, \nu > 0$$

where $\mu_i = \mathbb{E}(Y_i)$, ν is a shape parameter which quantifies the amount of overdispersion.
The variance of Y_i is $\mu_i + \mu_i^2 / \nu$. Clearly, the negative binomial distribution approaches a Poisson distribution when ν tends to ∞ (no overdispersion). Consequently, the ZINB distribution is given by

$$P(Y_i = y) \left\{ \begin{array}{ll} \pi_i + (1 - \pi_i)(1 + \frac{\mu_i}{\nu})^{-\nu} & y = 0 \\ (1 - \pi_i) \frac{\Gamma(y + \nu)}{\Gamma(\nu) y!} (1 + \frac{\mu_i}{\nu})^{-\nu} (1 + \frac{\nu}{\mu_i})^{-y}, & y = 1, 2, .. \end{array} \right.$$

The mean and variance of the ZINB distribution are $\mathbb{E}(Y_i) = (1 - \pi_i)\mu_i$ and $var(Y_i) = (1 - \pi_i)\mu_i(1 + \pi_i \mu_i + \mu_i / \mu)$, respectively. Observe that this distribution approaches the zero inflated Poisson distribution and the negative binomial distribution as $\nu \to \infty$ and $\pi_i \to 0$, respectively. If both $1/\nu$ and $\pi_i \approx 0$ then the ZINB distribution reduces to the Poisson distribution.
The ZINB regression model relates π_i and μ_i to covariates, that is,

$$\ln(\lambda_i) = \beta^T \mathbf{X}_i \quad \text{and} \quad \text{logit}(\pi_i) = \gamma^T \mathbf{W}_i, \quad i = 1, 2, .., n \tag{3}$$

where $\mathbf{X}_i$ and $\mathbf{W}_i$ are $p-$ and $q-$dimensional vectors of covariates pertaining to the ith subject, and with β and γ the corresponding vectors of regression coefficients, respectively.

3. APPLICATION

3.1. Data source

The data analyzed in this paper is a compulsory insurance for vehicle owners in Algeria. It is offered in the form of a package including obligatorily the civil liability guarantee and one or more other optional guarantees. This data set contains informations on 111086 Auto insurance policy reported during the period from Junuary through to December 2011.

3.2. Variables used in the data

PolicyID identifier of the insurance policy, **Numclaims** Number of claims,**Gender** gender of the client, **Age** age of the client, **AgeDL** age of the driving licence, **AgeVEH** age of the insured vehicle, **Brand** brand of the insured vehicle, **Power** fiscal power of the insured vehicle, **Duration** Coverage period of the insurance policy.

In this section, we did data cleaning and conducted exploratory analysis for the variables involved in this study. Histogram is used to display the distribution of the count response variable *numclaims*.

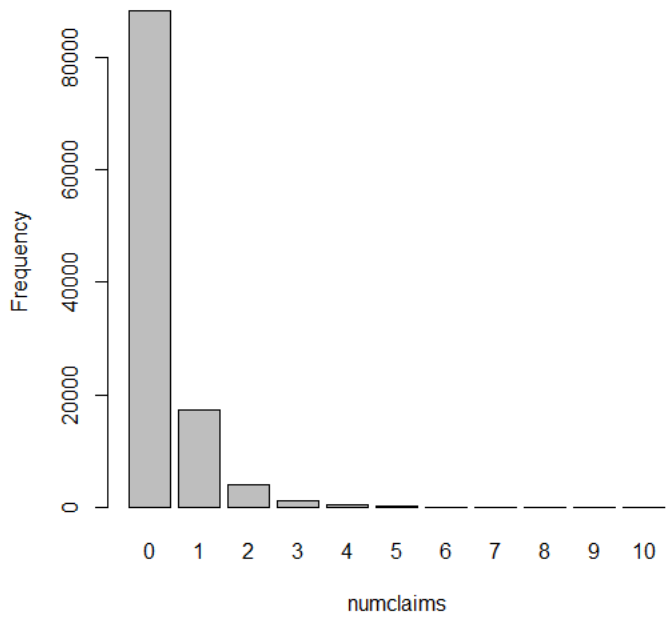

FIGURE 1 – Number of clamis made by clients

Figure1 shows histogram of the dependent variable.We clearly see a large number of zeros which cannot be modeled adequately with a Poisson model ; thus, the use of a zero-inflated count

model seems warranted.

3.2.1. Results

TABLE 1 – Estimated coefficients and standard errors for the count and zero Models.

parameter	variable	Poisson		NB		ZIP		ZINB	
		est.	se	est.	se	est.	se	est.	se
count									
β_1	Intercept	-3.399	0.045	-3.401	0.124	-3.083	0.067	-3.388	0.054
β_2	gender	-0.284	0.020	-0.294	0.024	-0.129	0.035	-0.215	0.027
β_3	agevr	0.615	0.032	0.599	0.036	0.827	0.052	0.673	0.037
β_4	agevi	0.144	0.036	0.142	0.039	0.146	0.038	0.141	0.039
β_5	ageDL	-0.028	0.003	-0.027	0.004	-0.026	0.004	-0.031	0.004
β_6	age25	-0.222	0.030	-0.223	0.035	-0.193	0.034	-0.195	0.036
β_7	age35	-0.167	0.033	-0.164	0.038	-0.182	0.036	-0.206	0.039
β_8	age45	-0.019	0.026	0.017	0.029	-0.052	0.029	-0.027	0.029
β_9	age55	0.052	0.016			0.062	0.026		
β_{10}	pow4	-0.427	0.112	-0.431	0.123			-0.414	0.123
β_{11}	brand3	0.095	0.014	0.097	0.017			0.079	0.017
β_{12}	brand4	-0.156	0.020	-0.153	0.022	-0.141	0.021	-0.150	0.022
β_{13}	brand5	0.204	0.015	0.205	0.017	0.088	0.025	0.182	0.017
β_{14}	age35 :ageDL	0.018	0.004	0.018	0.005	0.018	0.005	0.020	0.005
β_{15}	age45 :ageDL	0.009	0.003	0.008	0.003	0.010	0.003	0.010	0.003
θ	Theta			1.006	0.024			1.170	0.029
zero									
γ_1	Intercept					0.347	0.133	0.409	0.431
γ_2	gender					0.299	0.081	0.708	0.289
γ_3	agevr					0.568	0.110	1.295	0.266
γ_4	age45					-0.031	0.044		
γ_5	pow4					0.720	0.187		
γ_6	brand3					-0.196	0.033		
γ_7	brand5					-0.237	0.055		
γ_8	duration					-0.118	0.004	-0.525	0.042
AIC			144401.7		140705		140541.9		**139901.9**
BIC			144401.7		140848.9		140743.9		**140084.6**
log-ik			-72113.71		-70337.31		-70249.94		**-69931.94**
			(df=15)		(df=15)		(df=21)		**(df=19)**

In Table 1, we get outputs from four different models (Poisson, Negatif Binomial ,ZIP and ZINB). The first section of the output is for the positive-count process. The second section is for the zero-count process.In these outputs we post only the significant preditors (0.01 significance level). It can be seen that the parameter estimates of the Poisson,NB, ZIP and ZINB models are very similar for the count part. In addition, almost all these models have the same set of significant variables.
For purpose of comparison, we also report log-likelihood and AIC and BIC values at the bottom of the Table 1. The Poisson regression model had the largest criterion value, demonstrating the worst fit to the data. For the other three models, the ZINB model had smaller AIC and BIC values comparing with NB and ZIP models, while we find a little difference between NB and

ZIP criterion values. Among all the fitted models, the Zero-inflated Negative Binomial model had the smallest AIC and BIC values, so ZINB is the best choice for our response variable.

4. SIMULATION STUDY

We assess the performance of our proposed approach and contrast it to computing algorithms for penalized log-likelihood functions. The penalty functions include the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010).
Our approach is implemented in **R** (**R Development Core Team, 2012**), with the parameter estimates of the ZI model obtained from the **zeroinfl** function of package **pscl** (Jackman, 2011). Dr A. Buu kindly provided her R code (Buu et al., 2011) and we used the **R** package **mpath** (Wang et al., 2014b) to implement the approach of Wang et al. (2014a) and Wang et al. (2015). In both penalized approaches, we use the default parameter setting.

In this study, we consider one of the scenarios in Wang et al. (2015), namely their example 1. A ZINB model is considered with 20 covariates for each regression , so that $p = q = 20$. For three sample sizes $n = 500, 1000, 1500$, the predictors are randomly drawn from a $N_{20}(0,\Sigma)$ distribution, where Σ has elements $\rho^{|i-j|}$, for $i, j = 1,..,20$, with $\rho = 0.4$. The parameters are set to $\theta = 2$,

$$\beta = (1.10, 0, 0, 0, -0.36, 0, 0, 0, 0, 0, 0, 0, 0, -0.32, 0, 0, 0, 0, 0, 0, 0)$$

and

$$\gamma = (0.30, -0.48, 0, 0, 0, 0.4, 0, 0, 0, 0, 0.44, 0, 0.44, 0, 0, 0, 0, 0, 0, 0, 0)$$

We consider 1000 replications.

Results

TABLE 2 – **Simulation results with $n = 500$. Medians and standard deviations (in parentheses) of MSE, PE, $\hat{\theta}$, sensitivity and specificity.**

Method	MSE	Sensitivity	Specificity	PE	$\hat{\theta}$
		NB component			
ZINB-LASSO	0.074(0.02)	0.333(0.079)	1(0.032)	2.84(0.595)	1.113(0.625)
ZINB-MCP	0.071(0.017)	0.333(0.081)	1(0.22)	2.81(0.748)	1.225(0.445)
ZINB-SCAD	0.073(0.019)	0.333(0.098)	1(0.052)	2.825(0.626)	1.177(0.552)
Backward Elimination	0.095(0.02)	1(0.219)	0(0.314)	2.95(0.835)	1.463(0.566)
ZINB 1%	0.066(0.019)	0.333(0.094)	1(0.044)	2.77(0.613)	1.405(0.437)
ZINB 5%	0.067(0.02)	0.333(0.095)	1(0.043)	2.795(0.617)	1.395(0.459)
ZINB 10%	0.076(0.023)	0.333(0.103)	0.994(0.042)	2.83(0.627)	1.227(0.568)
		Zero component			
ZINB-LASSO	0.063(12872503.434)	0.143(0.132)	1(0.068)		
ZINB-MCP	609.299(69489682.931)	0.857(0.275)	0.464(0.403)		
ZINB-SCAD	0.08(33472586.102)	0.286(0.165)	1(0.103)		
Backward Elimination	95616.67(1045559.185)	1(0.238)	0(0.318)		
ZINB 1%	0.06(1.035)	0.286(0.116)	1(0.035)		
ZINB 5%	0.062(1.154)	0.286(0.114)	1(0.031)		
ZINB 10%	0.07(1.38)	0.143(0.083)	1(0.018)		

TABLE 3 – **Simulation results with $n = 1000$. Medians and standard deviations (in parentheses) of MSE, PE, $\hat{\theta}$, sensitivity and specificity.**

Method	MSE	Sensitivity	Specificity	PE	$\hat{\theta}$
		NB component			
ZINB-LASSO	0.074(0.017)	0.33(0.074)	1(0.032)	2.74(0.468)	1.117(0.45)
ZINB-MCP	0.072(0.014)	0.33(0.068)	1(0.218)	2.685(0.497)	1.2(0.379)
ZINB-SCAD	0.074(0.014)	0.33(0.01)	1(0.059)	2.717(0.475)	1.165(0.354)
Backward Elimination	0.092(0.016)	1(0.27)	0(0.39)	2.849(0.574)	1.074(0.286)
ZINB 1%	0.071(0.017)	0.33(0.099)	0.994(0.043)	2.705(0.473)	1.283(0.33)
ZINB 5%	0.068(0.014)	0.33(0.093)	1(0.043)	2.653(0.46)	1.34(0.298)
ZINB 10%	0.068(0.013)	0.33(0.093)	1(0.043)	2.655(0.461)	1.334(0.296)
		Zero component			
ZINB-LASSO	0.05(925263.184)	0.286(0.19)	1(0.064)		
ZINB-MCP	0.114(55177868.007)	0.857(0.227)	0.643(0.411)		
ZINB-SCAD	0.066(16026221.81)	0.286(0.187)	1(0.074)		
Backward Elimination	93009.101(8206888.167)	1(0.235)	0(0.387)		
ZINB 1%	0.049(1.1)	0.286(0.145)	1(0.022)		
ZINB 5%	0.046(0.787)	0.286(0.148)	1(0.035)		
ZINB 10%	0.047(0.69)	0.286(0.143)	1(0.036)		

TABLE 4 – **Simulation results with $n = 1500$. Medians and standard deviations (in parentheses) of MSE, PE, $\hat{\theta}$, sensitivity and specificity.**

Method	MSE	Sensitivity	Specificity	PE	$\hat{\theta}$
		NB component			
ZINB-LASSO	0.071(0.016)	0.333(0.074)	1(0.034)	2.68(0.426)	1.18(0.394)
ZINB-MCP	0.07(0.013)	0.333(0.063)	1(0.218)	2.7(0.453)	1.245(0.341)
ZINB-SCAD	0.074(0.012)	0.333(0.102)	1(0.067)	2.67(0.426)	1.159(0.296)
Backward Elimination	0.089(0.016)	1(0.306)	0.056(0.435)	2.875(0.508)	1.038(0.225)
ZINB 1%	0.07(0.013)	0.333(0.095)	1(0.042)	2.627(0.415)	1.313(0.244)
ZINB 5%	0.068(0.011)	0.333(0.097)	1(0.043)	2.62(0.414)	1.32(0.236)
ZINB 10%	0.069(0.011)	0.333(0.102)	1(0.045)	2.62(0.414)	1.31(0.236)
		Zero component			
ZINB-LASSO	0.045(40.817)	0.286(0.228)	1(0.059)		
ZINB-MCP	0.055(47933187.228)	0.857(0.2)	0.714(0.4)		
ZINB-SCAD	0.055(11422552.628)	0.286(0.216)	1(0.081)		
Backward Elimination	62558.811(1196548.234)	1(0.2)	0.071(0.438)		
ZINB 1%	0.04(0.628)	0.429(0.174)	1(0.025)		
ZINB 5%	0.038(0.387)	0.429(0.159)	1(0.035)		
ZINB 10%	0.038(0.39)	0.429(0.154)	1(0.036)		

5. REFERENCES

[1] AKAIKE H.[1973], *Information Theory and an Extension of the Maximum Likelihood Principle*
In Proceedings of International Symposium on Information Theory. B.N. Petrov et F. Czaki,Budapest.

[2] Cameron, AC, Trivedi, PK.[1998], *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.

[3] Cameron, AC, Trivedi, PK.[2005], *Microeconometrics : Methods and Applications*. Cambridge University Press, Cambridge.

[4] Diane Lambert.[1992], *Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing*. Technometrics, 34, 1-14.

[5] HILBE J.[2011], *Negative Binomial Regression,* Cambridge University Press, 2011.

[6] Mullahy, J.[1986], *Specification and Testing of Some Modified Count Data Models*. Journal of Econometrics, 33, 341-365.

[7] Nelder J.A., Wedderburn R.W.M.[1972], *Generalized Linear Models* . Journal of the Royal Statistical Society, Series A, 135, 370-384.

[8] Wang, Z., Ma, S., Wang, C.-Y., Zappitelli, M., Devarajan, P., and Parikh, C. [2014a], *EM for regularized zero-inflated regression models with applications to postoperative morbidity after cardiac surgery in children.*. Statistics in Medicine, 33, 5192–5208. ISSN 1097-0258.

[9] Wang Z, with contributions from Achim Zeileis, Jackman S, Ripley B et al. [2014b], *mpath : Regularized linear models. URL .R package version 0.1-17.*.

[10] Wang Z, Ma S and Wang CY [2015], *Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany*. Biometrical Journal, 57, 867–84.