Supplementary material to "Stochastic variable selection strategies for zero-inflated models"

Eva Cantoni [1] and Marie Auda [1]

[1] Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Geneva (Switzerland)

Address for correspondence: Eva Cantoni, Research Center for Statistics and Geneva School of Economics and Management, University of Geneva (Switzerland), 40, Bd du Pont d'Arve, 1211 GENEVA 4.

E-mail: Eva.Cantoni@unige.ch.

Phone: (+41) (0)22 379 8240.

Fax: (+41) (0)22 379 8299.

Abstract:

Key words: excess zeros; ZI model; hurdle model; variable selection; stochastic search.

1 Detailed specification of the ZIP and ZINB model

In this Section, we give the detailed expressions of model (1) of the main manuscript for the ZIP and ZINB model.

When $\mathcal{F}_{\mu_i,\boldsymbol{\theta}}$ is the Poisson distribution $\mathcal{P}(\mu_i)$, the probability mass function is

$$P_{\mathcal{F}_{\mu_i,\boldsymbol{\theta}}}(Y_i = y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!},$$

yielding

$$P(Y_i = y_i) \quad = \quad \begin{cases} \pi_i + \{1-\pi_i\}\exp(-\mu_i) & y_i = 0 \\ \{1-\pi_i\}\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} & y_i = 1, 2, \ldots \end{cases} \tag{1.1}$$

For this model, the expectation of Y_i is $E(Y_i) = (1-\pi_i)\mu_i$, and its variance is $Var(Y_i) = (1-\pi_i)(\mu_i + \pi_i\mu_i^2)$, showing the overdispersion naturally induced.

When $\mathcal{F}_{\mu_i,\boldsymbol{\theta}}$ is the negative binomial $NB(\mu_i, \theta)$, the probability mass function is

$$P_{\mathcal{F}_{\mu_i,\boldsymbol{\theta}}}(Y_i = y_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y_i+1)\Gamma(\theta)} \left(\frac{\mu_i}{\mu_i+\theta}\right)^{y_i} \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta},$$

where $1/\theta$ is the overdispersion parameter.

$$P(Y_i = y_i) \quad = \quad \begin{cases} \pi_i + \{1-\pi_i\}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta} & y_i = 0 \\ \{1-\pi_i\}\frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{y_i}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta} & y_i = 1, 2, \ldots \end{cases} \tag{1.2}$$

For this model, the expectation of Y_i is $E(Y_i) = (1-\pi_i)\mu_i$, and its variance is $Var(Y_i) = (1-\pi_i)(\mu_i + \mu_i^2/\theta + \pi_i\mu_i^2)$, showing again the overdispersion naturally induced.

2 Details on the operating transition kernel $q(\boldsymbol{\psi}^l|\boldsymbol{\upsilon}^{(s-1)})$

Denote by $\boldsymbol{p}^z = (p_1^z, \ldots, p_{p+q}^z)^T$ the $(p+q) \times 1$ vector containing the p-values of a z-test on the coefficients of each covariate in the full model (i.e. the model including all the available covariates). For a model $\boldsymbol{\upsilon}^{(s-1)}$ we denote the set of neighboring models $\boldsymbol{\psi}_{\boldsymbol{\upsilon}^{(s-1)}} = \{\boldsymbol{\psi}^1, \ldots, \ldots, \boldsymbol{\psi}^{p+q}\}$, where each neighboring model $\boldsymbol{\psi}^l = (\psi_1^l, \ldots, \psi_{p+q}^l)$ is such that $\sum_{i=1}^{p+q} |\boldsymbol{\upsilon}_i^{(s-1)} - \psi_i^l| = 1$, that is, each neighboring model differs from $\boldsymbol{\upsilon}^{(s-1)}$ in that it either includes an additional covariate or it excludes a present one. For each $\boldsymbol{\psi}^l \in \Psi_{\boldsymbol{\upsilon}^{(s-1)}}$, the transition kernel is defined by

$$q(\boldsymbol{\psi}^l|\boldsymbol{\upsilon}^{(s-1)}) = \frac{(1-p_l^z) \times E_l + p_l^z \times (1-E_l)}{\sum_{m=1}^{r} \{(1-p_m^z) \times E_m + p_m^z \times (1-E_m)\}},$$

where $E_l = 1$ if $\sum_{i=1}^{p+q} (\boldsymbol{\upsilon}_i^{(s-1)} - \psi_i^l) = 1$, that is if $\boldsymbol{\psi}^l$ includes an extra variable with respect to $\boldsymbol{\upsilon}^{(s-1)}$ and 0 otherwise.

To give an example let us consider a model $\boldsymbol{\upsilon}$ with $p = 2$ potential covariates to be included in the binary part and $q = 2$ others in the count part of the model, so that $p + q = 4$. Then, assume that the vector of p-values for the full ZI model $\boldsymbol{\upsilon}^{Full} = (1, 1, 1, 1)$ is $(0.2, 0.6, 0.1, 0.05)$. For the given submodel $\boldsymbol{\upsilon}^{(s-1)} = (1, 0, 1, 0)$ we obtain a set of $p + q = 4$ possible neighboring models

$$\Psi_{\boldsymbol{\upsilon}^{(s-1)}} = \begin{cases} \boldsymbol{\psi}^1 & = (0, 0, 1, 0), \\ \boldsymbol{\psi}^2 & = (1, 1, 1, 0), \\ \boldsymbol{\psi}^3 & = (1, 0, 0, 0), \\ \boldsymbol{\psi}^4 & = (1, 0, 1, 1). \end{cases}$$

The transition kernels for each neighboring model of $\Psi_{\boldsymbol{\upsilon}^{(s-1)}}$ are computed as follow

$$\begin{aligned} q(\boldsymbol{\psi}^1|\boldsymbol{\upsilon}^{(s-1)}) &= \frac{0.2}{[0.2+(1-0.6)+0.1+(1-0.05)]} = 0.12, \\ q(\boldsymbol{\psi}^2|\boldsymbol{\upsilon}^{(s-1)}) &= \frac{(1-0.6)}{[0.2+(1-0.6)+0.1+(1-0.05)]} = 0.24, \\ q(\boldsymbol{\psi}^3|\boldsymbol{\upsilon}^{(s-1)}) &= \frac{0.1}{[0.2+(1-0.6)+0.1+(1-0.05)]} = 0.06, \\ q(\boldsymbol{\psi}^4|\boldsymbol{\upsilon}^{(s-1)}) &= \frac{(1-0.05)}{[0.2+(1-0.6)+0.1+(1-0.05)]} = 0.58. \end{aligned}$$

The sum of the transition probabilities equals 1. It is more likely that $\boldsymbol{\psi}^4$ will be chosen (given that the p-value is small for last covariate which is not included in $\boldsymbol{\upsilon}^{(s-1)}$). The inverse kernels can be computed in a similar way.

3 Simulation Study 1

3.1 Regularization approaches

Regularization techniques allow to fit a model and simultaneously perform variable selection. The estimates are obtained maximizing the penalized likelihood

$$Q(\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}) = \ell(\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}) - n\sum_{j=2}^{p+1} p_{a_j}(|\beta_j|) - n\sum_{k=2}^{q+1} p_{b_k}(|\gamma_k|). \tag{3.1}$$

The penalties p_{a_j} and p_{b_k} can take several forms: either the LASSO penalty (**?**)) $p_\tau(\alpha) = \tau|\alpha|$ or the SCAD penalty (**?**))

$$p_\tau(|\alpha|) = \begin{cases} \tau|\alpha| & \text{if } 0 \le |\alpha| < \tau \\ -(|\alpha|^2 - 2c\tau|\alpha| + \tau^2)/[2(c-1)] & \text{if } \tau \le |\alpha| < c\tau \\ (c+1)\tau^2/2 & \text{if } |\alpha| \ge c\tau, \end{cases}$$

for $\tau \geq 0$ and $c > 2$, or the MCP penalty (**?**))

$$p_\tau(\alpha) = \tau \int_0^\alpha \left(1 - \frac{x}{c\tau}\right)_+ dx,$$

for $c > 1$ (in practice $c = 3.7$), and where $(u)_+$ denotes the positive part of u.

In **?**) the above approach is considered for the ZIP model with either the LASSO or SCAD penalty, whereas in **?**) and **?**) all three penalties are used for the ZIP and ZINB model respectively. These contributions pursue the same goal of optimizing the function (3.1), but they differ in their implementation. **?**) uses a local quadratic approximation of the log-likelihood function and a local linear approximation of the penalty functions, so that the LARS algorithm (**?**)) can be used. In **?**) and **?**) an EM-algorithm is used in conjunction with a coordinate descent algorithm. The two approaches also differ in the way they choose the smoothing parameters a_j and b_k. In **?**), three versions are considered: 1) use two tuning parameters $\tau_1 = a_j$ for all j and $\tau_2 = b_k$ for all k, treating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ separately, 2) use one tuning parameter τ and then set $a_j = \tau SE(\hat{\beta}_j^0)$ and $b_k = \tau SE(\hat{\gamma}_k^0)$, 3) use two tuning parameters τ_1 and τ_2 and set $a_j = \tau_1 SE(\hat{\beta}_j^0)$ and $b_k = \tau_2 SE(\hat{\gamma}_k^0)$, where SE stands for standard error and where $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\gamma}}^0$ are the unpenalized estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Then use BIC to choose the τ parameter (or τ_1, τ_2). In **?**) and **?**), the tuning parameters $a_j = \lambda_1$ for all j and $b_k = \lambda_2$ for all k are chosen on the basis of BIC on a grid between λ_{min} and $\lambda_{max} = \epsilon\lambda_{min}$ for several choices of ϵ.

Table 1: Median p-values (across the 300 simulated datasets) of a z-test on a ZIP full model (including all the available covariates) for each covariate included in the true data generating model from which the data were generated, according to the simulation design.

Setting	x_3	x_7	x_{10}	z_1	z_6	z_{10}
Independent ($n = 300$)	0.247	0.005	0.102	0.000	0.077	0.000
Correlated ($n = 300$)	0.229	0.008	0.229	0.000	0.079	0.002
Independent ($n = 600$)	0.116	0.000	0.021	0.000	0.012	0.000
Correlated ($n = 600$)	0.099	0.000	0.160	0.000	0.015	0.000

3.2 Study 1 setting

For the simulation setting of Study 1, Table 1 gives the p-values obtained from a z-test when we fit the full model (i.e. including all the available variables). It illustrate the range of significance levels and henceforth the range of signal strength, for each variable.

3.3 Additional results for $n = 300$

In the following, we provide additional results pertaining to simulation Study 1 of the main paper. They are briefly discussed in the paper itself.

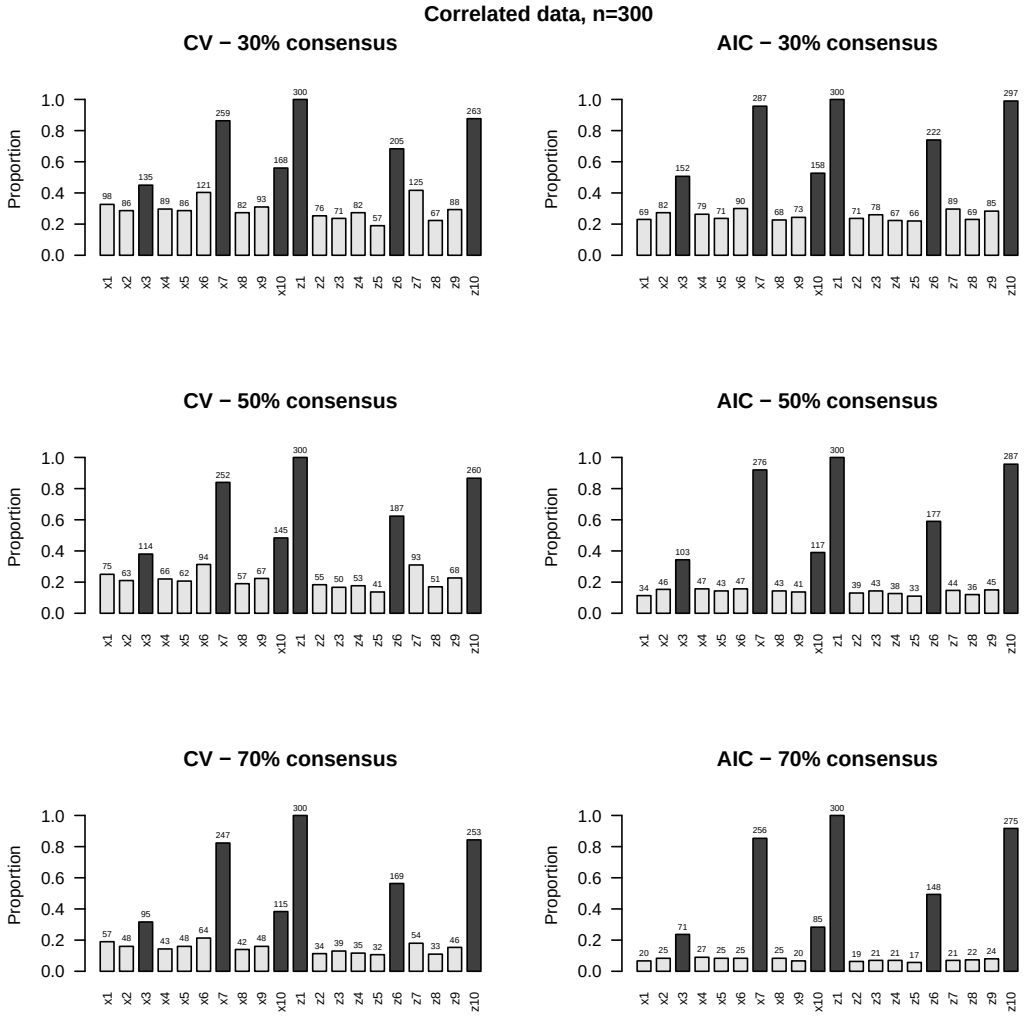

Figure 1: Proportion of simulated samples (over the last $5,000$ visited models of each MCMC chain) including the variable in the consensus model. In dark grey are the variables included in the true data generating model.

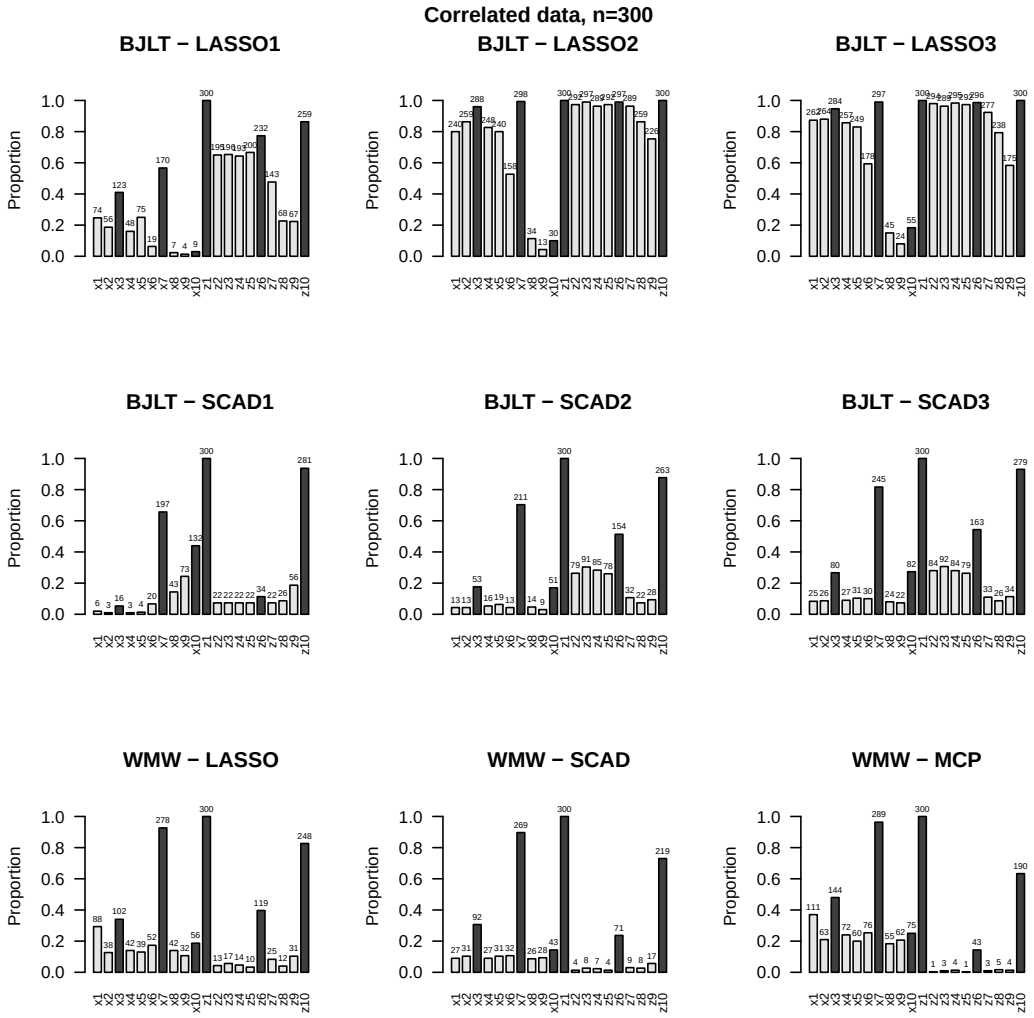

Figure 2: Proportion of simulated samples including the variable in final model. In dark grey are the variables included in the true data generating model.

Table 2: Summary of number of votes for the best model and summary of number of models visited, burn-in period removed ($n = 300$).

	CV			AIC		
	Independent data					
	1st Q	Median	3rd Q	1st Q	Median	3rd Q
number of votes for best model	419.20	687.00	920.50	79.75	126.00	172.60
number of models visited	95.75	175.50	239.90	627.80	794.00	831.20
	Correlated data					
	1st Q	Median	3rd Q	1st Q	Median	3rd Q
number of votes for best model	341.80	669.00	851.10	66.00	109.00	146.00
number of models visited	107.00	187.50	255.70	652.00	864.00	897.50

Table 3: Number of times the true model has been correctly identified (Exact), the number of times one or more data generating variables are missing from the selected model, but with possibly many extra spurious variables (Underfit) and the number of times a largest model containing the true one (Overfit) has been selected ($n = 300$).

	Correlated data						
	Underfit (≤ 3)	Underfit (-2)	Underfit (-1)	Exact	Overfit ($+1$)	Overfit ($+2$)	Overfit (> 3)
CV - 30%	49	97	118	1	0	3	32
CV - 50%	66	112	100	2	1	7	12
CV - 70%	89	124	74	3	2	3	5
AIC - 30%	16	108	119	0	7	18	32
AIC - 50%	66	123	90	5	10	4	2
AIC - 70%	111	127	58	3	1	0	0
min CV	125	112	56	2	2	1	2
min AIC	144	132	24	0	0	0	0
votes CV	69	108	99	1	4	5	14
votes AIC	36	129	114	0	12	12	3
BJLT-LASSO1	133	94	68	0	0	0	5
BJLT-LASSO2	0	14	259	0	0	0	27
BJLT-LASSO3	0	17	234	0	0	0	49
BJLT-SCAD1	175	103	22	0	0	0	0
BJLT-SCAD2	152	101	41	0	1	1	4
BJLT-SCAD3	107	119	63	1	0	2	8
WMW-LASSO	129	99	58	1	3	2	8
WMW-SCAD	173	83	36	1	3	2	2
WMW-MCP	154	93	49	0	0	1	3

3.4 Results for $n = 600$

In this section, we provide and discuss the results for $n = 600$ for Study 1.

Globally speaking all the effects observed when $n = 300$ show up more strongly when $n = 600$. More precisely, in terms of marginal frequency of appearance, the true signal is better identified and the superfluous variables more clearly discarded (smaller marginal frequency of appearance), see Table 4. The results on the consensus model (Figures 3 and 4) show the same general pattern as for $n = 300$, but with global better performance for each approach, as one would expect. The same comment holds for the penalized approaches presented in Figures 5 and 6. For the results based on the minimum criterion (Table 5), we see that a larger sample size improves the performance in recognizing the data generating signals for both CV and AIC, but the proportion of inclusion of the superfluous variables is not much improved with respect to $n = 300$. Table 7 indicates that CV and AIC behave differently as a function of the samplesize: while with larger samplesize, AIC visits less models and the number of visits for the best model increases, the contrary is observed for CV.

When evaluating the performance in terms of model identification (Tables 8 and 9), we see that with $n = 600$ we have globally, for all techniques, less "heavy underfitting" (3 or more variable missing). This is particularly true for the 70% consensus model with both CV and AIC, the minimum criterion with AIC, BJLT-LASSO1, BJLT-SCAD (all three versions) and WMW-LASSO. On the down side, this has the consequence to increase overfitting.

Table 4: Summary of the marginal frequency of appearance computed over the 10'000 visited models of the Markov chain, $n = 600$. The variables included in the true model are highlighted in bold.

	Independent data						Correlated data					
	Binary part						Binary part					
	CV prediction error			AIC			CV prediction error			AIC		
	1st Q	Median	3rd Q	1st Q	Median	3rd Q	1st Q	Median	3rd Q	1st Q	Median	3rd Q
x_1	0.004	0.041	0.453	0.077	0.120	0.220	0.005	0.047	0.378	0.084	0.125	0.202
x_2	0.013	0.093	0.526	0.090	0.134	0.266	0.008	0.057	0.377	0.090	0.127	0.251
x_3	**0.029**	**0.579**	**0.986**	**0.179**	**0.496**	**0.873**	**0.082**	**0.715**	**0.994**	**0.219**	**0.566**	**0.928**
x_4	0.020	0.110	0.462	0.085	0.134	0.249	0.014	0.094	0.433	0.087	0.135	0.291
x_5	0.011	0.079	0.410	0.085	0.130	0.262	0.018	0.141	0.565	0.093	0.148	0.254
x_6	0.061	0.226	0.567	0.079	0.118	0.267	0.068	0.212	0.563	0.090	0.144	0.267
x_7	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**1.000**	**1.000**
x_8	0.011	0.058	0.279	0.079	0.112	0.224	0.015	0.088	0.422	0.087	0.135	0.241
x_9	0.007	0.049	0.331	0.080	0.121	0.232	0.011	0.074	0.277	0.094	0.142	0.265
x_{10}	**0.503**	**0.981**	**1.000**	**0.528**	**0.924**	**0.993**	**0.076**	**0.558**	**0.960**	**0.167**	**0.421**	**0.803**
	Count part						Count part					
	CV prediction error			AIC			CV prediction error			AIC		
	1st Q	Median	3rd Q	1st Q	Median	3rd Q	1st Q	Median	3rd Q	1st Q	Median	3rd Q
z_1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
z_2	0.009	0.069	0.351	0.079	0.122	0.256	0.010	0.050	0.362	0.092	0.137	0.267
z_3	0.014	0.108	0.516	0.086	0.129	0.231	0.014	0.118	0.438	0.088	0.134	0.270
z_4	0.015	0.080	0.427	0.082	0.124	0.245	0.008	0.063	0.375	0.089	0.150	0.272
z_5	0.006	0.052	0.410	0.077	0.118	0.243	0.003	0.039	0.288	0.083	0.127	0.234
z_6	**0.785**	**0.998**	**1.000**	**0.709**	**0.968**	**0.998**	**0.753**	**0.989**	**1.000**	**0.690**	**0.958**	**0.998**
z_7	0.076	0.268	0.645	0.084	0.127	0.262	0.075	0.249	0.626	0.092	0.144	0.281
z_8	0.006	0.064	0.446	0.079	0.136	0.265	0.009	0.053	0.406	0.088	0.134	0.259
z_9	0.008	0.066	0.448	0.078	0.128	0.270	0.006	0.045	0.318	0.091	0.132	0.232
z_{10}	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

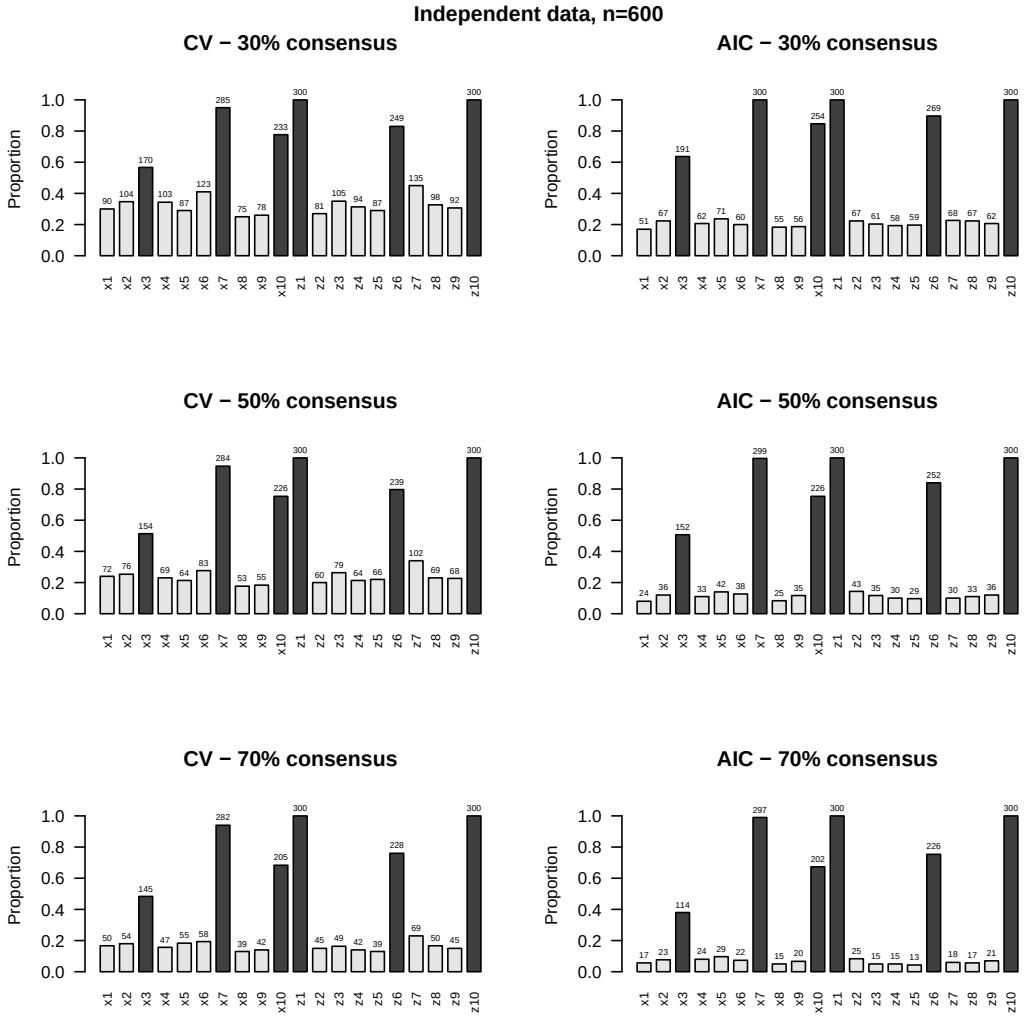

Figure 3: Proportion of simulated samples (over the last 5,000 visited models of each MCMC chain) including the variable in the consensus model. In dark grey are the variables included in the true data generating model.

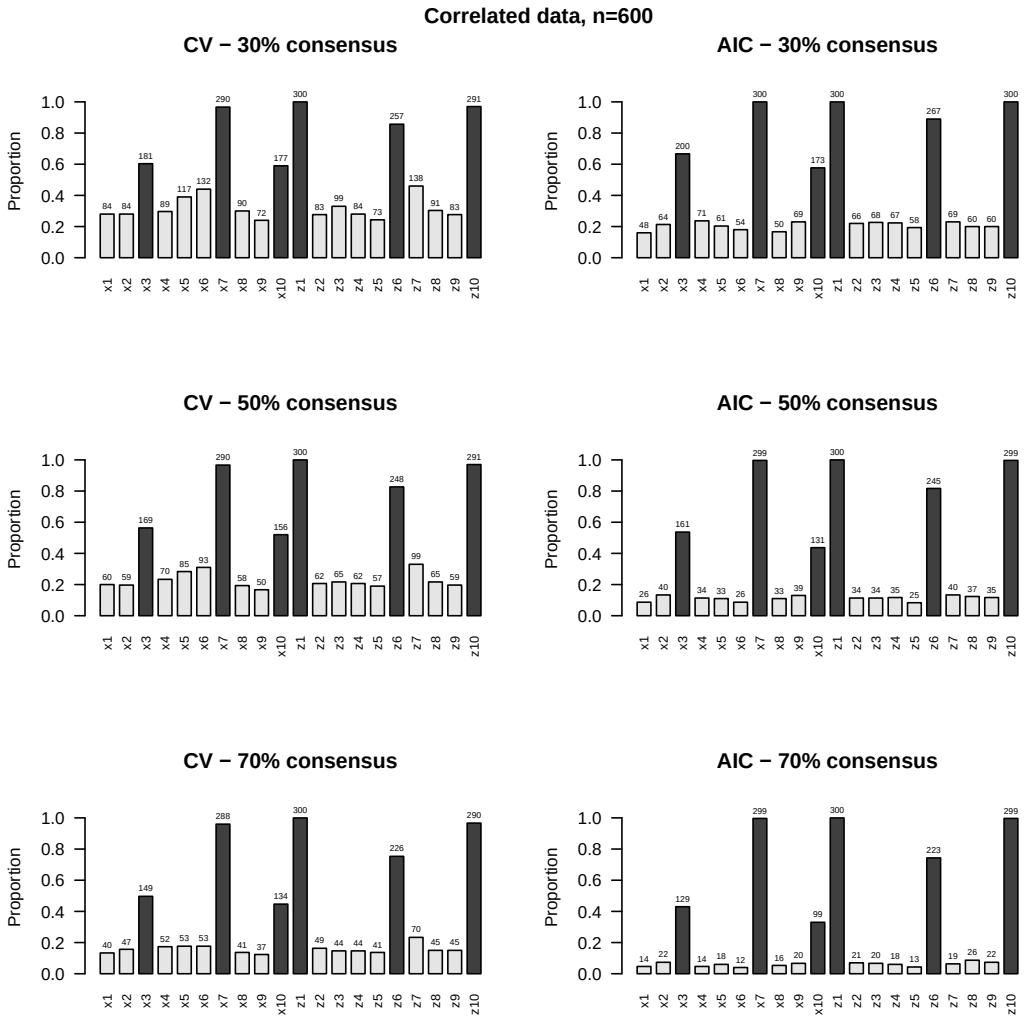

Figure 4: Proportion of simulated samples (over the last 5,000 visited models of each MCMC chain) including the variable in the consensus model. In dark grey are the variables included in the true data generating model.

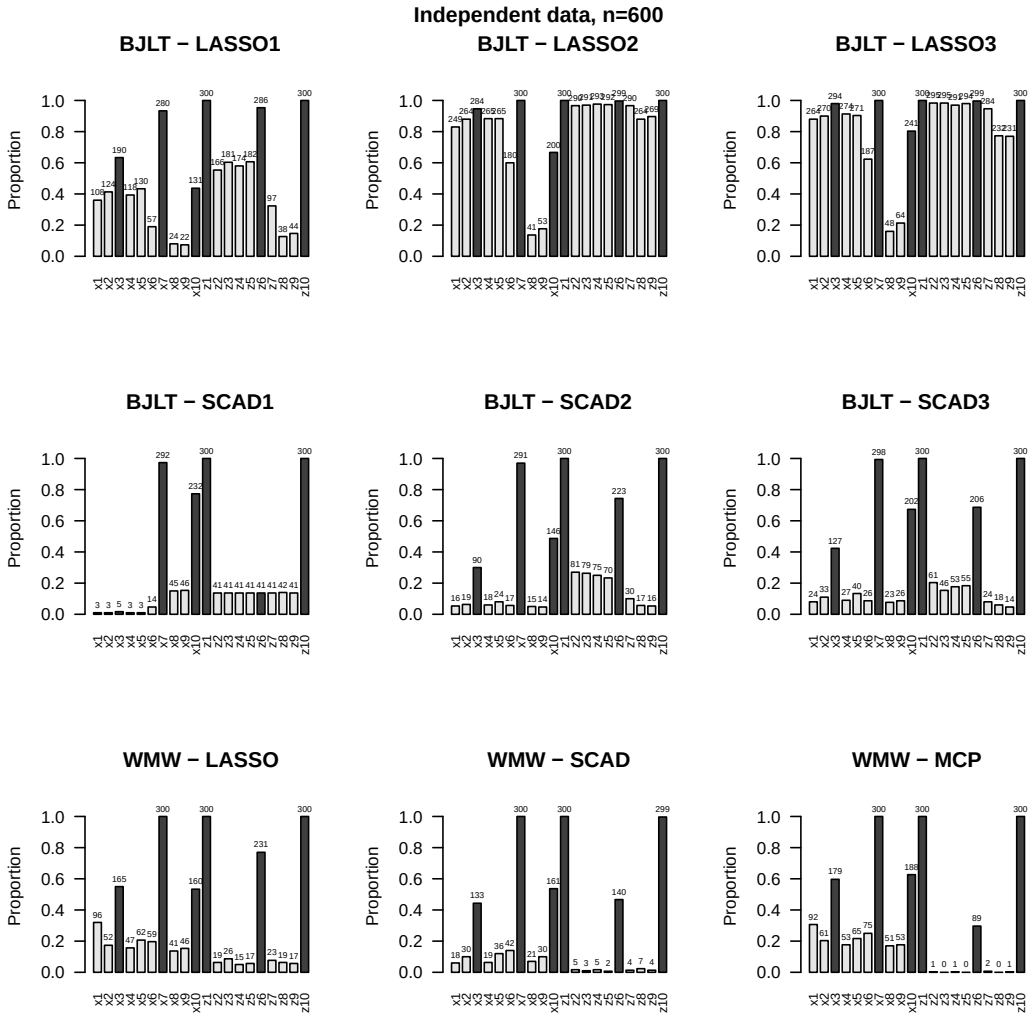

Figure 5: Proportion of simulated samples including the variable in final model. In dark grey are the variables included in the true data generating model.

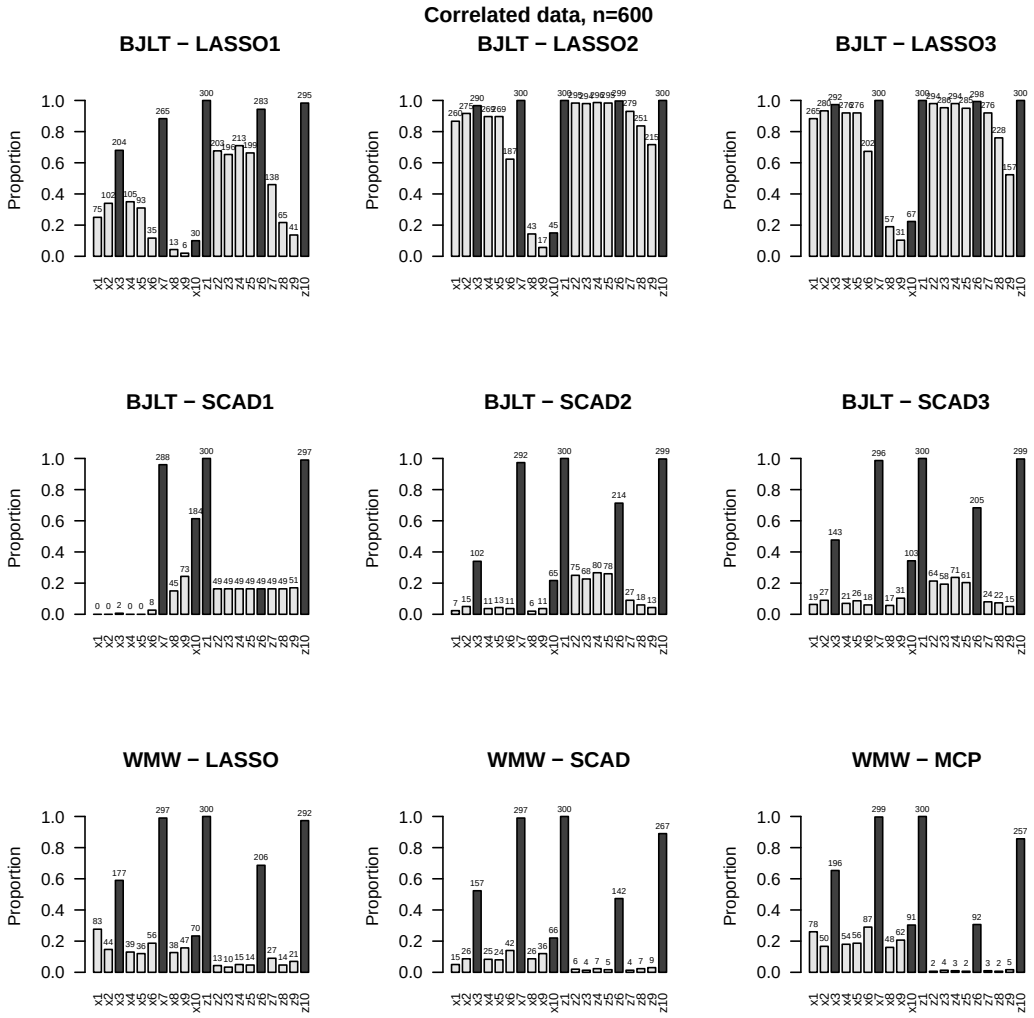

Figure 6: Proportion of simulated samples including the variable in final model. In dark grey are the variables included in the true data generating model.

Table 5: Proportion of simulated samples for which each variable is included in the optimal model that minimize the criterion across the entire MCMC chain ($n = 600$).

	Independent data						Correlated data				
	Binary part			Count part			Binary part			Count part	
	CV	AIC		CV	AIC		CV	AIC		CV	AIC
x_1	0.177	0.030	z_1	**1.000**	**1.000**	x_1	0.150	0.027	z_1	**1.000**	**1.000**
x_2	0.167	0.033	z_2	0.130	0.040	x_2	0.130	0.027	z_2	0.153	0.040
x_3	**0.430**	**0.243**	z_3	0.180	0.033	x_3	**0.420**	**0.293**	z_3	0.163	0.043
x_4	0.173	0.037	z_4	0.130	0.033	x_4	0.137	0.023	z_4	0.147	0.023
x_5	0.163	0.040	z_5	0.163	0.020	x_5	0.187	0.023	z_5	0.147	0.020
x_6	0.200	0.043	z_6	**0.693**	**0.627**	x_6	0.180	0.030	z_6	**0.640**	**0.630**
x_7	**0.900**	**0.983**	z_7	0.250	0.027	x_7	**0.930**	**0.980**	z_7	0.217	0.023
x_8	0.107	0.023	z_8	0.180	0.040	x_8	0.117	0.020	z_8	0.170	0.047
x_9	0.107	0.023	z_9	0.160	0.030	x_9	0.093	0.037	z_9	0.137	0.023
x_{10}	**0.603**	**0.510**	z_{10}	**1.000**	**1.000**	x_{10}	**0.397**	**0.203**	z_{10}	**0.957**	**0.997**

Table 6: Proportion of simulated samples for which each variable is included in the optimal model with largest number of votes over the last $5,000$ visited models of each MCMC chain ($n = 600$).

	Independent data						Correlated data				
	Binary part			Count part			Binary part			Count part	
	CV	AIC		CV	AIC		CV	AIC		CV	AIC
x_1	0.236	0.118	z_1	**1.000**	**1.000**	x_1	0.210	0.123	z_1	**1.000**	**1.000**
x_2	0.262	0.154	z_2	0.203	0.180	x_2	0.210	0.173	z_2	0.207	0.156
x_3	**0.522**	**0.587**	z_3	0.276	0.161	x_3	**0.560**	**0.591**	z_3	0.223	0.173
x_4	0.249	0.148	z_4	0.203	0.125	x_4	0.230	0.179	z_4	0.190	0.193
x_5	0.223	0.187	z_5	0.209	0.144	x_5	0.287	0.153	z_5	0.187	0.123
x_6	0.276	0.161	z_6	**0.794**	**0.872**	x_6	0.300	0.126	z_6	**0.833**	**0.834**
x_7	**0.950**	**1.000**	z_7	0.329	0.167	x_7	**0.963**	**0.997**	z_7	0.327	0.186
x_8	0.183	0.134	z_8	0.236	0.154	x_8	0.207	0.140	z_8	0.220	0.146
x_9	0.189	0.151	z_9	0.216	0.157	x_9	0.173	0.169	z_9	0.207	0.153
x_{10}	**0.741**	**0.800**	z_{10}	**1.000**	**1.000**	x_{10}	**0.513**	**0.515**	z_{10}	**0.970**	**1.000**

Table 7: Summary of number of votes for the best model and summary of number of models visited, burn-in period removed ($n = 600$).

Independent data						
	CV			AIC		
	1st Q	Median	3rd Q	1st Q	Median	3rd Q
number of votes for best model	359.00	635.00	770.30	128.00	207.00	285.30
number of models visited	134.00	204.00	274.90	449.00	578.00	588.80
Correlated data						
	CV			AIC		
number of votes for best model	334.2	596.5	755.5	107.0	165.0	216.2
number of models visited	133.50	243.50	302.50	535.00	662.00	680.00

Table 8: Number of times the true model has been correctly identified (Exact), the number of times one or more data generating variables are missing from the selected model, but with possibly many extra spurious variables (Underfit) and the number of times a largest model containing the true one (Overfit) has been selected ($n = 600$).

	Independent data						
	Underfit (≤ 3)	Underfit (-2)	Underfit (-1)	Exact	Overfit ($+1$)	Overfit ($+2$)	Overfit (> 3)
CV - 30%	6	48	149	1	1	9	86
CV - 50%	9	58	154	2	11	16	50
CV - 70%	12	74	155	6	11	19	23
AIC - 30%	0	22	142	1	23	44	68
AIC - 50%	3	49	164	16	35	24	9
AIC - 70%	12	86	153	19	21	8	1
min CV	27	95	140	3	9	16	10
min AIC	26	150	113	9	2	0	0
votes CV	10	57	154	2	8	14	55
votes AIC	1	31	160	5	42	45	20
BJLT-LASSO1	22	73	100	0	1	2	102
BJLT-LASSO2	0	6	105	0	0	0	189
BJLT-LASSO3	0	3	60	0	0	0	237
BJLT-SCAD1	64	199	37	0	0	0	0
BJLT-SCAD2	39	106	117	3	8	11	16
BJLT-SCAD3	18	92	129	9	16	19	17
WMW-LASSO	24	75	122	11	15	19	34
WMW-SCAD	58	92	108	16	13	10	3
WMW-MCP	45	97	115	6	10	15	12

Table 9: Number of times the true model has been correctly identified (Exact), the number of times one or more data generating variables are missing from the selected model, but with possibly many extra spurious variables (Underfit) and the number of times a largest model containing the true one (Overfit) has been selected ($n = 600$).

	Correlated data						
	Underfit (≤ 3)	Underfit (-2)	Underfit (-1)	Exact	Overfit ($+1$)	Overfit ($+2$)	Overfit (> 3)
CV - 30%	12	67	132	0	6	11	72
CV - 50%	16	83	130	6	12	15	38
CV - 70%	27	103	123	9	10	9	19
AIC - 30%	2	44	166	4	11	31	42
AIC - 50%	10	90	155	12	18	11	4
AIC - 70%	24	126	127	16	6	1	0
min CV	50	122	98	3	14	9	4
min AIC	46	177	75	2	0	0	0
votes CV	16	82	133	5	12	14	38
cites AIC	3	72	167	6	26	24	3
BJLT-LASSO1	37	75	160	0	0	0	28
BJLT ASSO2	0	9	248	0	0	0	43
BJLT-LASSO3	0	7	229	0	0	0	64
BJLT-SCAD1	100	172	28	0	0	0	0
BJLT-SCAD2	57	131	91	3	4	6	8
BJLT-SCAD3	36	117	112	6	9	10	10
WMW-LASSO	44	104	111	6	16	4	15
WMW-SCAD	89	101	75	5	9	9	12
WMW-MCP	78	112	84	9	8	4	5

4 Variable description of the docvisits data set.

In this section, we give some additional data description for the dataset used in Section 4 of our main article, namely an histogram of the number of visits to the doctor (Figure 7) and the variable description (Table 10).

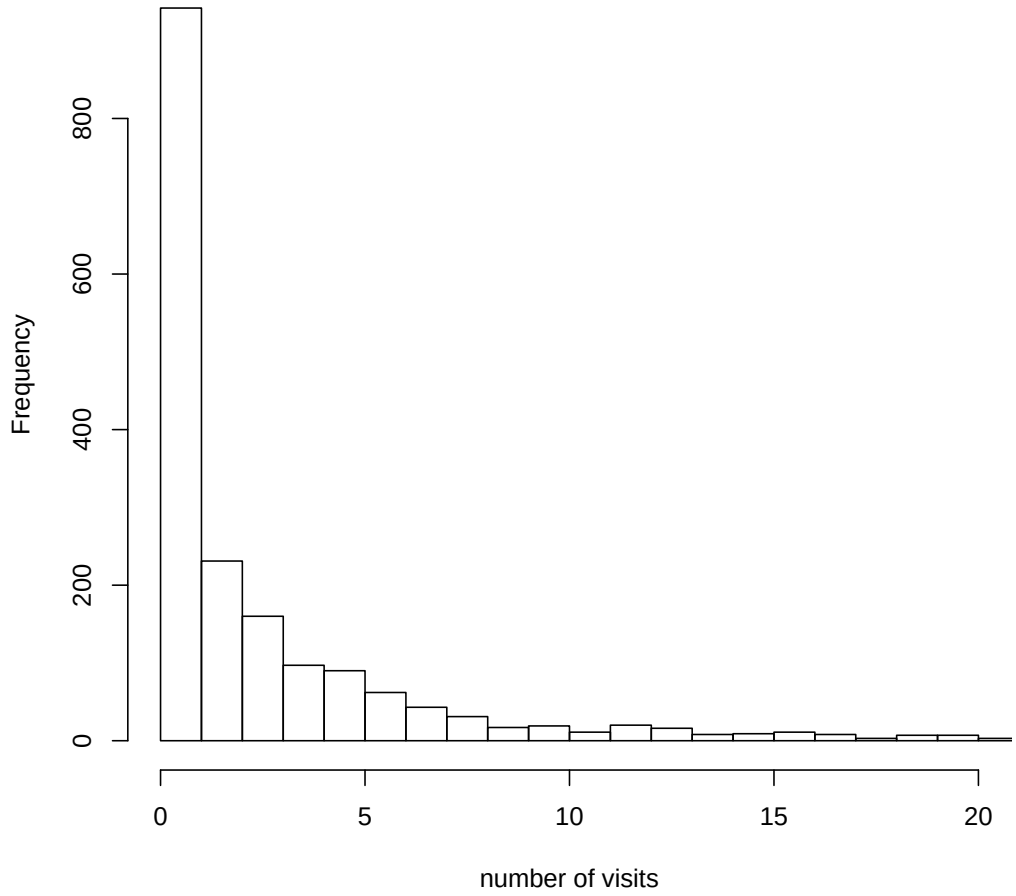

Figure 7: Histogram of the number of visits to the doctor for the docvisits data set.

Table 10: Description of the docvisits data set.

Variable	Description
docvisits	number of doctor visits in last 3 months
age	age
handicap	1 if handicapped, 0 otherwise
hdegree	degree of handicap in percentage points
married	1 if married, 0 otherwise
schooling	years of schooling
hhincome	household monthly net income, in German marks / 1000
children	1 if children under 16 in the household, 0 otherwise
self	1 if self employed, 0 otherwise
civil	1 if civil servant, 0 otherwise
bluec	1 if blue collar employee, 0 otherwise
employed	1 if employed, 0 otherwise
public	1 if public health insurance, 0 otherwise
addon	1 if add-on insurance, 0 otherwise
age30	1 if age ≥ 30
age35	1 if age ≥ 35
age40	1 if age ≥ 40
age45	1 if age ≥ 45
age50	1 if age ≥ 50
age55	1 if age ≥ 55
age60	1 if age ≥ 60

References

Buu, A., Johnson, N. J., Li, R., and Tan, X. (2011). New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in medicine*, **30**(18), 2326–2340.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, **32**(2), 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(3), 273–282.

Wang, Z., Ma, S., Wang, C.-Y., Zappitelli, M., Devarajan, P., and Parikh, C. (2014). EM for regularized zero-inflated regression models with applications to postoperative morbidity after cardiac surgery in children. *Statistics in Medicine*, **33**(29), 5192–5208. ISSN 1097-0258.

Wang, Z., Ma, S., and Wang, C.-Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. *Biometrical Journal*, **57**(5), 867–884.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942.