Bayesian Model Selection Methodology for Road Safety

Bahar Dadashova*, Blanca Arenas, José Mira, Francisco Aparicio

University Institute of Automobile Research (INSIA), Technical University of Madrid (UPM)
José Gutiérrez Abascal 2, 28006 Madrid, Spain

Abstract

Road accidents are a very relevant issue in many countries and macroeconomic models are very frequently applied by academia and administrations to reduce their frequency and consequences. The selection of explanatory variables and response transformation parameter within the Bayesian framework for the selection of the set of explanatory variables a TIM and 3IM (two input and three input models) procedures are proposed. The procedure also uses the DIC and pseudo - R^2 goodness of fit criteria. The model to which the methodology is applied is a dynamic regression model with Box-Cox transformation (BCT) for the explanatory variables and autorgressive (AR) structure for the response. The initial set of 22 explanatory variables are identified. The effects of these factors on the fatal accident frequency in Spain, during 2000-2012, are estimated. The dependent variable is constructed considering the stochastic trend component.

1. Data

The monthly data for fatal accidents in Spain during 2000-2012 are considered. There were initially 22 explanatory variables considered among which the following variables were selected (Table 1).

2. Model selection methodology

The model selection methodology mainly consists of three stages: 1)variable selection through TIM (two input model) and 3IM (three input model); 2) optimal BCT selection or differenecing of dependent variable; 3) model selection (Dadashova et al. , 2014).

2.1. Variable selection

$$Y_t = \sum_{k=1}^{K} \beta_k X_k^{\lambda_{X_i}} + u_t \tag{2.1}$$

$$K = \begin{cases} 2, \text{ if TIM} \\ 3, \text{ if 3IM} \end{cases} \tag{2.2}$$

$$i = \{1, 2, 3\} \tag{2.3}$$

$$u_t = \sum_{l=1}^{2} \rho_l u_{t-l} + w_t \tag{2.4}$$

The variable selection is carried out in two directions. First we estimate two-input models - TIM ($K = 2$ in equations (2.1)-(2.4)). Given that there are 22 variables the possible combinations are $\binom{22}{2} = 231$, thus 231 TIMs are estimated. The TIM variables are power transformed using 3 candidate BCT values $\lambda_X = (-0.5, 0.1, 0.5)$. Sice the variables in the same TIM are transformed with the same values of λ_X, there are $231 \cdot 3 = 693$, i.e. 3 sets of 231 TIMs. In the second procedure 3 input models - 3IMs are considered instead of two inputs ($K = 3$ in equations (2.1)-(2.4)). In this stage the variable combinations are $\binom{22}{3} = 1540$. Considering three candidate values of λ_X a total of $1540 \cdot 3 = 4620$, i.e. 3 sets of 1540 3IMs were defined.

The first 50 TIMs/3IMs with the highest R^2 from 3 sets of 231/1540 models are selected ($50 \cdot 3 = 150$ models in total). The final variables selected as the result of this stage are obtained the TIM/3IMs that coincide accross 3 sets of 50 models with the highest R^2 (Dadashova et al. , 2014).

The estimation methodology is MCMC which was implemented using R and WinBUGS softwares. The Gibbs sampler was run in $10,000$ iterations in 3 chains. Initially uninformative priors are considered: i.e. multivaraite normal for β_k and ρ_l, inverted gamma for σ_w^2 and uniform prior for y_0 (Chib , 1993).

2.2. Dependent variable

The optimal form of dependent variable is defined with respect to the independent variables. For this purpose there are 2 possibilities. First the dependent variable is power transformed with respect to $X_{k,\{k=1,8\}}$ (Venables and Ripley , 2002). The second option is the differencing of the dependent variable. For this purpose the hyperparameter Y_0 is assigned a uniform prior and estimated through MCMC. The results show that the goodness of fit measure and prediction accuracy measures in case of BC transformed dependent variable are better.

2.3. Model selection

The final models were estimated using the inputs selected through TIM and BC transformed dependent variable. The model structure is the same as depicted in the

*Corresponding author: bahar@etsii.upm.es

equations (2.1)-(2.4) where $K = \{1, 2, ..., 8\}$. In this case the variables belonging to the same group are transformed with the same value of λ_X (Gaudry and Lassarre , 2000). The selected variables initially belong to 6 groups (*G1, G2, G3, G4, G5, G6* (Table 1) in case of TIM variables and *G1, G2, G3, G4, G6, G7* in case of 3IM variables) where the last group has an only dummy variable. Therefore only the variables of previous groups are transformed, i.e. variables belonging to 5 groups. The candidate values for λ_X remain the same, i.e. $\lambda_X = (-0.5, 0.1, 0.5)$. Considering the number of groups (= 5) and candidate λ_X (= 3), a total of $3^5 = 243$ structural explanatory (SE) models are constructed and estimated. Since the model estimation is carried out in parallel the whole procedure in sections 2.1-2.3 is applied 3 times which we call scenarios: Scenario 1) TIM BCT; Scenario 2) TIM DIFF; and Scenario 3) 3IM (Table 1). In each 3 cases the model with better goodness of fit measures and the correct sign of the variables among the rest of the models.

3. Prediction analysis

The final model estimation is cross validated through prediction. For this purpose the model estimation was carried out using first 144 observations (2000-2011) and cross validated using the remaining 12 observations (2012). The prediction results for each of the 3 selected models are compared using 3 prediction accuracy measures: mean squared error (MSE), mean absolute error (MAE) and mean absolute prediction error (MAPE).

4. Conclusions

The variables selected through the methodology are very relevant for the road safety analysis. The results of the dependent variable show that the power transformation of the dependent variable can resolve the problem of stochastic trend present in the data and in fact the differencing of the data beforehand can result in overestimation.

References

Box, G. E., and Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26(2), 211 − 252.

Chib, S. (1993). Bayes regression with autoregressive errors: A Gibbs sampling approach. Journal of Econometrics, 58(3), 275 − 294.

Dadashova, B., Arenas, B., Mira, J. and Aparicio, F. (2014). Analysis of fatal road accidents in Spain. Model selection methodology. Under review.

Gaudry, M. and Lassarre, S. (2000). Structural road accident models. The international DRAG family. Oxford: Elsevier Science.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (2002). Bayesian variable and transformation selection in linear regression. Journal of Computational and Graphical Statistics, 11(3), 485 − 507.

Venables, W. N., and Ripley, B. D. (2002). Random and mixed effects. In Modern Applied Statistics With S (271 − 300). Springer New York.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. J. Wiley and Sons, New York.

Table 1: ESTIMATION RESULTS.

Variable Estimates										**Hyperparameters**			**Prediction Accuracy**			**Goodness of fit**		
Group	Definition	Variable	SE99		[a] SE109		[b] SE215		[c]		SE99	SE109	SE215		SE99	SE109	SE215	
			η_X [d]	S.D.	η_X	S.D.	η_X	S.D.										
G1 [e]	Vehicle km traveled	VKM	0.4	0.017	3.59	228	0.63	0.28		ρ_1	0.19	-0.33	0.49	MAE	0.06	0.1	0.08	
	Diesel consumption	CONOIL					0.49	0.69		ρ_2	0.22	-0.02	0.46	MAPE	15.07	23.28	17.19	R^2
G2 [f]	Total unemployment	PARO	-0.19	0.028	-0.93	263				τ	1.08	0.01	1.51	MSE	372	960	490	*DIC*
	Industrial production index	IPI	0.09	0.104	0.11	167	-0.87			y_0		344						
G3 [g]	Young drivers	COND2					0.6	0.38		λ_{X1}	0.5	-0.5	0.1					
	License suspended	SUSP	-0.44	0.139	-1.13	276	-0.75	0.25		λ_{X2}	0.5	-0.5	0.5					
	Radar checks	RADAR	-0.65	0.366	-1.97	292				λ_{X3}	0.1	-0.5	0.5					
G4 [h]	Older vehicles (10 yrs)	VEH10	0.22	8.803	0.21	42	0.97	1.58		λ_{X4}	-0.5	0.1	0.1					
G5 [i]	High capacity roads (%)	LONRAC	-1.6	3.473	-0.12	63				λ_{X5}	0.1	0.1	0.5					
G6 [j]	Penalty Point System	PPS	-0.03	0.576	-0.001	12	-0.02	0.74		λ_Y	0.45		0.4					
G7 [k]	Rainfall	PREC					0.01	0.01										

[a]Selected from Scenario 1) TIM BCT : variables selected from TIMs and dependent variable is BC transformed
[b]Selected from Scenario 2) TIM DIFF: variables selected from TIMs and dependent variable is differenced
[c]Selected from Scenario 3) 3IM: variables selected from 3IMs and dependent variable is BC transformed
[d]Elasticity
[e]Exposure
[f]Economic factors
[g]Driver characteristics and surveillance
[h]Vehicle characteristics
[i]Road infrastructure
[j]Legislative measures
[k]Weather conditions