Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 1 (2015), pp. 295–306

http://www.ripublication.com/gjpam.htm

Wald Test and Distance-Based Generalized Linear Models. Actuarial Application

Eva Boj and Teresa Costa

Departament de Matemàtica Econòmica,
Financera i Actuarial, Universitat de Barcelona,
Avinguda Diagonal 690, 08034 Barcelona, Spain.
E-mail: evaboj@ub.edu, tcosta@ub.edu

Abstract

The Distance-Based Generalized Linear Model (DB-GLM) is a generalization of the classical GLM to the DB framework. The DB-GLM is non-linear on original predictors because its information is entered in the model by means of a squared distances matrix. In a previous work, we defined influence coefficients for original factors in the DB-GLM to measure its importance. Now, with the aim to test the null hypothesis that each coefficient is equal to a fixed value, we define a *t*-like test statistic, we estimate its null hypothesis distribution by a bootstrapping pairs methodology and use it to obtain percentile confidence intervals. We make an example with actuarial data, and we fit the models with the `dbstats` package for `R`.

AMS subject classification:
Keywords: Distance-based generalized linear model, Influence coefficients, Wald test, Confidence intervals, Actuarial science, `R`.

1. Introduction

The DB-GLM, defined in Boj et al. [2], extends the ordinary GLM (McCullagh and Nelder [14] allowing information on predictors to be entered as interdistances between observation pairs instead of as individual coordinates. In turn, these interdistances may have been computed from arbitrary, non-numerical observed predictors.

The estimation process of a DB-GLM is schematically as follows: a Euclidean configuration is obtained by a metric multidimensional scaling-like procedure, then the linear predictor of the underlying GLM is a linear combination of the resulting Euclidean coordinates, latent variables in the model. Therefore influence coefficients of the original observed predictors cannot be computed as in the ordinary GLM.

In Boj et al. [4] we proposed a definition of local influence coefficients for the DB-GLM depending on the nature of risk factors (numerical or categorical/binary). These coefficients measure the relative importance of each observed variable. In this paper, we study how to adapt the Wald test of predictor significance to the DB-GLM environment.

To this end, firstly we apply the definition of influence coefficients and the bootstrap by pairs methodology to estimate the distribution of coefficients, as is given in Boj et al. [4]. In this way we are able to estimate the coefficients of the DB-GLM and its associated standard errors. Then, we propose a procedure to adapt the Wald statistic to the DB-GLM. We construct *simple confidence intervals* by using a standard normal distribution, and *percentile t confidence intervals* by using a bootstrap t^* distribution, where both types of confidence intervals are understood in the sense defined in MacKinnon [12]. The t^* distribution of the percentile intervals follows the null hypothesis of the test in the bootstrap data generation process. In this way, the percentile t confidence intervals are useful to test the null hypothesis that a coefficient is equal to a fixed real value.

We illustrate the calculation of percentile t confidence intervals with a well known actuarial dataset. We estimate the related DB-GLM by using the dbglm function of the `dbstats` `R` package (Boj et al. [3], R Development Core Team [15]).

The paper is structured as follows. In Section 2 we describe the proposed procedure for the Wald test in the DB-GLM. Firstly, in Sub-section 2.1, we recall the definition of influence coefficients; secondly, in Sub-section 2.2, we construct simple and percentile t confidence intervals. In Section 3 we make an example. Finally, in Section 4, we conclude.

2. Hypothesis testing: the Wald test

The main objective of this work is to adapt the Wald test to the DB-GLM. The Wald test contrasts the null hypothesis $H_0 : \beta_j = \beta_0$. The statistic:

$$\tau_j = \frac{\hat{\beta}_j - \beta_0}{std\left(\hat{\beta}_j\right)} \tag{2.1}$$

follows (in the classical GLM) an asymptotically t distribution. In (2.1) $\hat{\beta}_j$ is the unrestricted estimate of the parameter $\hat{\beta}_j$ that is being tested and $std(\hat{\beta}_j)$ is its standard error.

In the next two subsections, we propose a procedure to estimate a bootstrapped t^* distribution for the DB-GLM that follows the null hypothesis, $H_0 : \beta_j = \beta_0$. First we recall the definition of influence coefficients and the bootstrap by pairs methodology to estimate coefficient distribution explained in Boj et al. [4]. Then, we show how to construct simple and percentile confidence intervals, taking into account the bootstrapped t^* distribution that follows the null hypothesis of the test.

2.1. Influence coefficients for the DB-GLM

Assume we have a response variable Y observed for n individuals, and a set of p risk factors F_j for $j = 1, \ldots, p$. In ordinary GLM the relation between response and linear predictor is

$$\hat{y} = g^{-1}\left(\hat{\eta}\right) = g^{-1}\left(\hat{\beta}_0 + F_1 \cdot \hat{\beta}_1 + F_2 \cdot \hat{\beta}_2 + \cdots + F_p \cdot \hat{\beta}_p\right),$$

where we can measure the influence of the predictor F_j by means of $\hat{\beta}_j$ for $j = 1, \ldots, p$. But in DB-GLM the relation of each observable predictor F_j with the prediction is not linear. The idea underlying our definition in Boj et al. [4] of influence is to mimic that of the $\hat{\beta}_j$ in GLM.

The influence of each F_j we want to quantify depends on a reference/virtual individual $\mathbf{f}^0$ which we take as reference or origin. Let $\mathbf{f}^0 = \left(f_1^0, f_2^0, \ldots, f_p^0\right)$ be the vector with the p predictor values of a reference individual. For instance $\mathbf{f}^0$ consists of mean or median in numerical coordinates and the mode in binary or qualitative coordinates.

For categorical (or binary) predictors we defined the influence coefficients $\hat{\beta}_j$ for $j = 1, \ldots, p$ as the increment in the estimated linear predictor $\hat{\eta}$ when the j-th predictor value of $\mathbf{f}^0$ changes to another level. As a short notation:

$$\beta_j = \Delta j\ \hat{\eta}\big|_{\mathbf{f}^0} \text{ for } j = 1, \ldots, p$$

For quantitative predictors we define the influence coefficients $\hat{\beta}_j$ for $j = 1, \ldots, p$ as:

$$\beta_j = \left.\frac{\partial \hat{\eta}}{\partial F_j}\right|_{\mathbf{f}0} \text{ for} j = 1, \ldots, p$$

the speed of the estimated linear predictor $\hat{\eta}$ changes as $\mathbf{f}^0$ moves along the curve:

$$\mathbf{f}0 + t \times s_j \left(0, \ldots, 0, \underset{j-th}{1}, 0, \ldots, 0\right), t \in (-\varepsilon, +\varepsilon)$$

where s_j is the standard deviation of the j-th quantitative predictor.

2.2. Bootstrap confidence intervals

There is an extensive literature on the numerous ways to construct bootstrap confidence intervals. MacKinnon [12] proposed that the simplest approach is to calculate the bootstrap standard error and use it to construct confidence intervals based on the normal distribution. A simple confidence interval for β_j at level $1 - \alpha$ is:

$$\left[\hat{\beta}_j - std^*\left(\hat{\beta}_j\right) \times z_{1-}\alpha_{/2}, \hat{\beta}_j + std^*\left(\hat{\beta}_j\right) \times z_{1-}\alpha_{/2}\right], \tag{2.2}$$

where $z_{1-}\alpha_{/2}$ denotes the $1 - {}^{\alpha}/_{2}$ quantile of the standard normal distribution. If, e.g., $\alpha = 0.05$ this is equal to 1.96. The simple bootstrap confidence interval can be modified

so that it will be centred on a bias-corrected estimate of β_j by simply replacing $\hat{\beta}_j$ in (2.2) by

$$\tilde{\beta}_j = \hat{\beta}_j - \left(\bar{\hat{\beta}}_j^* - \hat{\beta}_j\right) = 2\hat{\beta}_j - \bar{\hat{\beta}}_j^*.$$

In Boj et al. [4] we proposed a bootstrap by pairs methodology to estimate the distribution of β_j and its standard deviation, $std^*\left(\hat{\beta}_j\right)$, where we refer for a detailed description of the method.

The pairs bootstrap is easy to implement and it can be applied to a wide range of models. The resampling technique consists of n response-predictor pairs from the original data, see Boj et al. [1], [4], Davidson and Hinkley [5], Efron and Tibshirani [6] or Hall [9] among other references. Then, one can generate B bootstrap samples from which to estimate the statistic of interest. However, the bootstrap data generation process in the bootstrap by pairs does not impose any restrictions on β_j.

If we are testing some restrictions, e.g. $H_0 : \beta_j = \beta_0$, we need to modify the bootstrap data generation process in such a way that the given restrictions are enforced, yielding a valid bootstrap test statistic. A way to proceed is to use the modified bootstrap test statistic:

$$\hat{\tau}_j^b = \frac{\hat{\beta}_j^b - \hat{\beta}_j}{std^*\left(\hat{\beta}_j\right)} \tag{2.3}$$

where $\hat{\beta}_j^b$ is the estimate of β_j from the b-th bootstrap sample, for $b = 1, \ldots, B$ where B is the sample size, and the denominator $std^*\left(\hat{\beta}_j\right)$ is the standard error of the $\hat{\beta}_j$ distribution. As the estimate of β_j from the bootstrap samples should, on average, be equal to $\hat{\beta}_j$, the null hypothesis tested by $\hat{\tau}_j^b$ is true in the pairs bootstrap data generation process. In this way, we can compare the statistic of the original sample,

$$\hat{\tau}_j = \frac{\hat{\beta}_j - \beta_0}{std^*\left(\hat{\beta}_j\right)},$$

with the bootstrap t^* distribution given by (2.3), and calculate a p-value by:

$$\hat{p}^*\left(\hat{\tau}_j\right) = \frac{1}{B}\sum_{b=1}^{B} I\left(\left|\hat{\tau}_j^b\right| > \left|\hat{\tau}_j\right|\right),$$

or by,

$$\hat{p}^*\left(\hat{\tau}_j\right) = 2\min\left(\frac{1}{B}\sum_{b=1}^{B} I\left(\hat{\tau}_j^b \leq \hat{\tau}_j\right), \frac{1}{B}\sum_{b=1}^{B} I\left(\hat{\tau}_j^b > \hat{\tau}_j\right)\right).$$

An interval that has better properties than the simple bootstrap confidence interval is the percentile t confidence interval. A percentile t confidence interval for β_j at level

$1-\alpha$ is defined as:

$$\left[\hat{\beta}_j - std^*\left(\hat{\beta}_j\right) \times t^*_{1-\alpha/2}, \hat{\beta}_j - std^*\left(\hat{\beta}_j\right) \times t^*_{\alpha/2}\right] \tag{2.4}$$

where t^*_δ is the δ quantile of the bootstrap distribution of the t^* statistic defined in (2.3). For example, if $\alpha = 0.05$ then $\alpha/2 = 0.025$ and $1 - \alpha/2 = 0.975$ and $t^*_{\alpha/2}$ is the 0.025 quantile of the bootstrap t^* distribution and $t^*_{1-\alpha/2}$ is the 0.975 one. The t^* distribution given by (2.3) follows the null hypothesis, then the percentile t confidence interval (2.4) is usefully for the hypothesis testing $H_0 : \beta_j = \beta_0$ with β_0 a fixed real value. With these type of confidence intervals it is not necessary to repeat the calculations if we want to change the value of β_0, unlike what happens with p-values.

3. Application

We use a data set on Swedish third-party motor insurance in 1977 described in Hallin and Ingenbleek [10] and also used in Boj et al. [2], [4]. Data are included in the package `faraway` for `R` under the name *motorins*. The total number of observations is $n = 295$ corresponding to different non-empty risk groups. We analyze claim frequency, defined by the number of claims and the exposure variable number of insured in policy-years. There are three risk factors: Distance (kilometers travelled by year, with 5 levels), Bonus (level in the scale of Bonus, with continuous numerical values from 1 to 7) and Make (with 9 nominal categories). We code Bonus and Distance (using its class mark) as numerical variables and Make as categorical nominal. We assume a Poisson error distribution and the logarithmic link.

We fit DB-GLM to the main effects of the three risk factors, using the dbglm function in the `dbstats` package (see Boj et al. [3]. The similarity is computed with Gower's similarity index (`metric = "gower"`), taking into account all the geometric variability (`rel.gvar = 1`), i.e., the model named `dbglm1` in the Appendix A of Boj et al. [2], with a residual deviance of 454.1 on 276 degrees of freedom:

```
Call:
dbglm(formula = yfactor(MakeC) + KmC + BonC , data = Motor1,
family = poisson(link = "log"), method = "rel.gvar", metric
= "gower",
weights = w, rel.gvar = 1)

family:  poisson
metric:  gower

Degrees of Freedom:  294 Total (i.e.  Null); 276 Residual 236
Null Deviance:  6978 237
Residual Deviance:  454.1
```

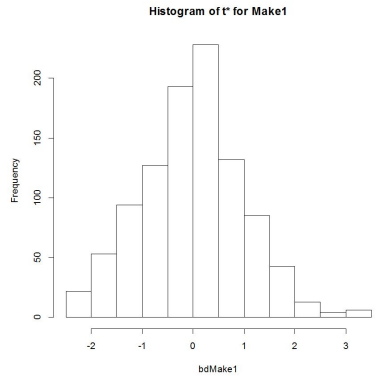

Figure 1: Bootstrap t^* distribution of $\hat{\beta}_{Make1}$ for the `dbglm1` model.

In Boj et al. [4] eleven coefficients were estimated: nine for the levels *Make1* to *Make9*, and two for the numerical factors *Distance* and *Bonus*. The linear predictor was:

$$\begin{aligned}&\eta = \beta_0 + F_1 \cdot \beta_1 + F_2 \cdot \beta_2 + \cdots + F_{10} \cdot \beta_{10} + \varepsilon = \\ &\beta_{Make1} + F_{Make2} \cdot \beta_{Make2} + F_{Make3} \cdot \beta_{Make3} + F_{Make4} \cdot \beta_{Make4} + \\ &F_{Make5} \cdot \beta_{Make5} + F_{Make6} \cdot \beta_{Make6} + F_{Make7} \cdot \beta_{Make7} + \\ &F_{Make8} \cdot \beta_{Make8} + F_{Make9} \cdot \beta_{Make9} + F_{Km} \cdot \beta_{Km} + F_{Bon} \cdot \beta_{Bon} + \varepsilon.\end{aligned}$$

The reference individual chosen was:

$$\mathbf{f}^0 = \left(Make = 1, Km = \overline{Km}, Bon = \overline{Bon}\right) = (1, 9683.82, 5.58)\,,$$

being the class *Make* = 1 the corresponding to the intercept term β_0.

In Boj et al. [4] one can find, in Table 1, the results of the estimated influence coefficients for this model, `dbglm1`, and the corresponding standard errors using size $B = 1000$ for the bootstrap. Table 2 Table 2 contains simple bootstrap confidence intervals, (2.2), at level 0.95.

Now, we complete the example with a quantitative measure for the significance of predictors using the Wald test. We construct the corresponding percentile t confidence intervals, (2.4), at level 0.95. In Figures from 1 to 11 we show the histograms of the bootstrap t^* distributions given by (2.3) for the different predictors of the `dbglm1` model. In Table 3, one can find: in the first column the estimated betas; in the second column the 97.5 quantile of the bootstrapped null distribution given by the 1000 values of (2.3); in the third column the 2.5 quantile of the same bootstrap distribution; and in the fourth column the percentile t confidence intervals given by (2.4). As a result, we do not have the 0 value in any of the percentile t confidence intervals of Table 1, and it means that all the coefficients are significant in the `dbglm1` model.

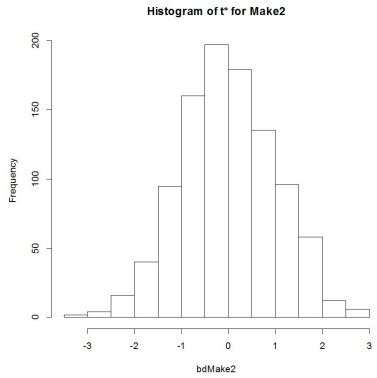

Figure 2: Bootstrap t^* distribution of $\hat{\beta}_{Make2}$ for the `dbglm1` model.

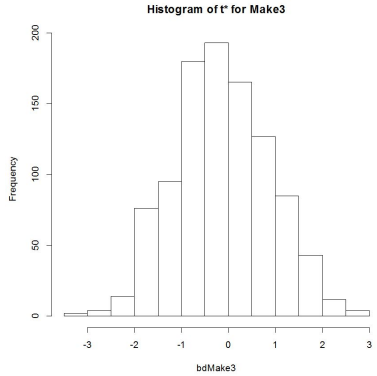

Figure 3: Bootstrap t^* distribution of $\hat{\beta}_{Make3}$ for the `dbglm1` model.

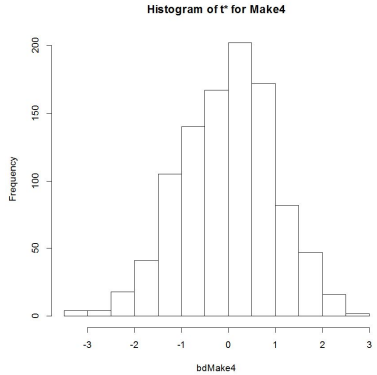

Figure 4: Bootstrap t^* distribution of $\hat{\beta}_{Make4}$ for the `dbglm1` model.

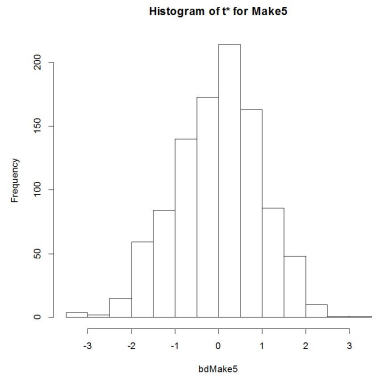

Figure 5: Bootstrap t^* distribution of $\hat{\beta}_{Make5}$ for the `dbglm1` model.

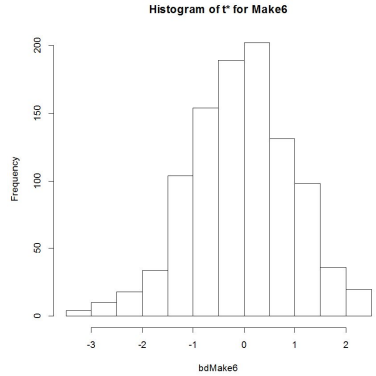

Figure 6: Bootstrap t^* distribution of $\hat{\beta}_{Make6}$ for the `dbglm1` model.

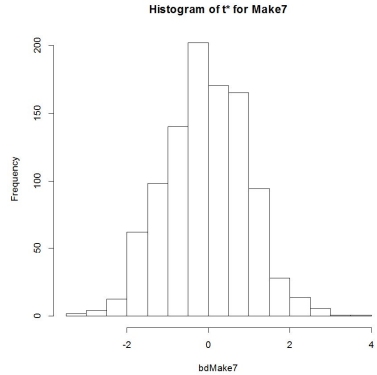

Figure 7: Bootstrap t^* distribution of $\hat{\beta}_{Make7}$ for the `dbglm1` model.

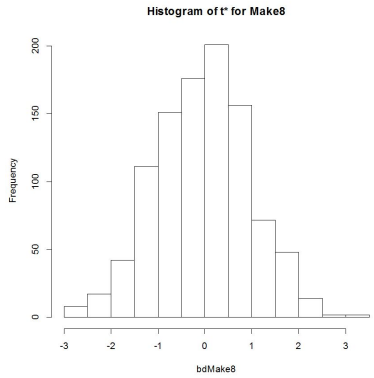

Figure 8: Bootstrap t^* distribution of $\hat{\beta}_{Make8}$ for the `dbglm1` model.

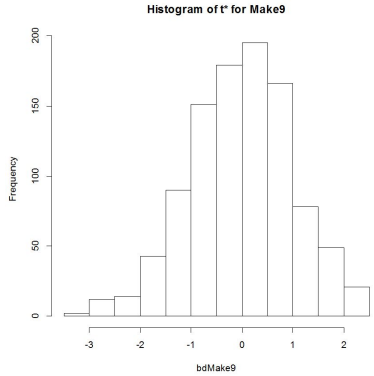

Figure 9: Bootstrap t^* distribution of $\hat{\beta}_{Make9}$ for the `dbglm1` model.

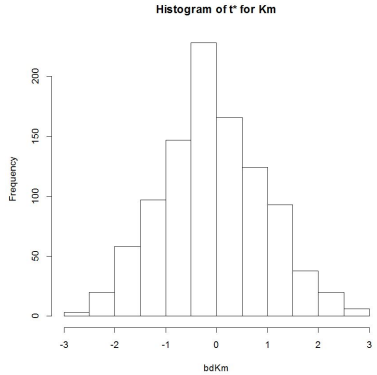

Figure 10: Bootstrap t^* distribution of $\hat{\beta}_{Km}$ for the `dbglm1` model.

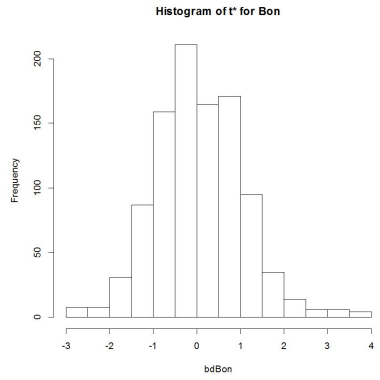

Figure 11: Bootstrap t^* distribution of $\hat{\beta}_{Bon}$ for the `dbglm1` model.

Table 1: Estimated coefficients; quantiles 97.5 and 2.5 of the t^* distributions; and percentile t confidence intervals assuming $\alpha = 0.05$, for the `dbglm1` model.

	$\hat{\beta}$	$t^*_{1-\alpha/2}$	$t^*_{\alpha/2}$	Percentile t confidence intervals
$\hat{\beta}_{Make1}$	$-1.857e+0$	1.962086	-1.908919	$[-1.860e+0, -1.853e+0]$
$\hat{\beta}_{Make2}$	$1.312e-01$	1.882888	-1.849325	$[1.251e-01, 1.371e-01]$
$\hat{\beta}_{Make3}$	$-2.142e-01$	1.828144	-1.976043	$[-2.260e-01, -2.014e-01]$
$\hat{\beta}_{Make4}$	$-4.977e-01$	1.930592	-2.144194	$[-5.028e-01, -4.919e-01]$
$\hat{\beta}_{Make5}$	$1.239e-01$	1.81118	-1.968902	$[1.178e-01, 1.305e-01]$
$\hat{\beta}_{Make6}$	$-3.880e-01$	1.916472	-2.03158	$[-3.926e-01, -3.832e-01]$
$\hat{\beta}_{Make7}$	$-1.304e-01$	1.968073	-1.909681	$[-1.410e-01, -1.200e-01]$
$\hat{\beta}_{Make8}$	$1.363e-01$	1.872734	-1.998162	$[1.081e-01, 1.664e-01]$
$\hat{\beta}_{Make9}$	$-2.361e-02$	1.799623	-2.038099	$[-2.597-02, -2.093e-02]$
$\hat{\beta}_{Km}$	$1.069e-05$	2.022189	-1.953856	$[1.050e-05, 1.087e-05]$
$\hat{\beta}_{Bon}$	$-3.733e-02$	2.113138	-1.685741	$[-4.111e-02, -3.431e-02]$

4. Conclusions

In Boj et al. [4] we define -local valid- influence coefficients for the DB-GLM. Additionally, we propose a bootstrap methodology to estimate standard errors and to calculate simple bootstrap confidence intervals as an informative measure. The proposed bootstrap methodology is bootstrapping pairs which could be adequate when we use DB regression models (see Boj et al. [1]).

The pairs bootstrap is very easy to implement and it can be applied to a wide range of models. However it sufers from two major deficiencies (see MacKinnon [11], [12],[13]). The first is that the bootstrap data generation process does not impose any restriction on

β_j. Then, if we are testing restrictions on β_j, as opposed to estimating standard errors or forming simple confidence intervals, we need to modify the bootstrap test statistic so that it will test something that will be true in the bootstrap data generation process. Or, alternatively, we can modify the resampling scheme so that the null hypothesis will be respected in the bootstrap data generating process, see Boj et al. [1] and Flachaire [7], [8]. The other deficiency of the pairs bootstrap is that, compared with the residual bootstrap (when it is valid) and with the wild bootstrap, the pairs bootstrap generally does not yield very accurate results. But the pairs bootstrap is less sensible to the hypotheses of the model than the residual bootstrap. And the estimated standard error via the pairs bootstrap offers reasonable results when some hypotheses of the model are not satisfied.

To complete the study of influence coefficients for the DB-GLM initiated in Boj et al. [4], in this work we propose a procedure to obtain an estimation of the Wald statistic. Our objective is to contrast the null hypothesis, $H_0 : \beta_j = \beta_0$. For this aim, we compute percentile t confidence intervals given by formula (2.4). The bootstrapped t^* distribution given by (2.3) follows the null hypothesis, then the percentile t confidence interval (2.4) is adequate for the hypothesis testing $H_0 : \beta_j = \beta_0$ given β_0 a fixed real value. With these type of confidence intervals it is not necessary to repeat the calculations if we want to contrast the null hypothesis for different values of β_0, unlike what happens when using p-values.

In the example, we have calculated percentile t confidence intervals for the coefficients of the *motorins* dataset related to the model with the main effects, named tt dbglm1. And we have obtained that all risk factors could be entered as a tariff variables in the final model.

The most important result of this work is that with the defined bootstrap percentile t confidence intervals we can study in an statistical way the significance of the influence coefficients defined in Boj et al. [4] for the DB-GLM. And this is an important question in actuarial rate-making, where the selection of risk factors is the basis of the problem.

Acknowledgments

Work supported in part by the Spanish Ministerio de Educación y Ciencia and FEDER, grant MTM2010-17323, and by Generalitat de Catalunya, AGAUR grant 2014SGR152.

References

[1] E. Boj, M.M. Claramunt and J. Fortiana, Selection of predictors in distance-based regression, *Communications in Statistics A. Theory and Methods*, 36, 87–98, 2007.

[2] E. Boj, P. Delicado, J. Fortiana, A. Esteve and A. Caballé. Local distance-based generalized linear models using the dbstats package for R. *Documentos de Trabajo de la Xarxa de Referència en Economia Aplicada (XREAP)*, XREAP2012-11, 2012.

[3] E. Boj, A. Caballé, P. Delicado, and J. Fortiana, *dbstats: Distance-Based Statistics (dbstats)*. R package version 1.0.3, 2013, URL http://CRAN.R-project.org/package=dbstats.

[4] E. Boj, T. Costa, J. Fortiana and A. Esteve, Assessing the Importance of Risk Factors in Distance-Based Generalized Linear Models. *Methodology and Computing in Applied Probability*, http://link.springer.com/article/10.1007/s11009-014-9415-6, 2014.

[5] A.C. Davidson and D.V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, New York, 1997.

[6] B. Efron and J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1998.

[7] E. Flachaire, A better way to bootstrap pairs. *Economics Letters*, 64, 257–262, 1999.

[8] E. Flachaire, Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, 49, 361–376, 2005.

[9] P. Hall, *The Bootstrap and Edgeworth Expansion. Springer Series in Statistics*, Springer, New York, 1992.

[10] M. Hallin and J.F. Ingenbleek, The Swedish automobile portfolio in 1977, A statistical study, *Scandinavian Actuarial Journal*, 49–64, 1983.

[11] J.G. MacKinnon, Bootstrap inference in econometrics, *The Canadian Journal of Economics* 35, 4, 615–645, 2002.

[12] J.G. MacKinnon, Bootstrap methods in econometrics, *The Economic Record* 82, special issue, september 2006, s2–s18, 2006.

[13] J.G. MacKinnon, Bootstrap hypothesis testing, *Queen's Economics Department Working Paper*, Number 1127, 2007.

[14] P. McCullagh and J.A. Nelder, *Generalized Linear Models (2nd ed)*., Chapman and Hall, London, 1989.

[15] R Development Core Team (2014), R*: A Language and Environment for Statistical Computing*. Vienna, Austria, URL http://www.R-project.org/.