DOCUMENTS DE TREBALL

DE LA FACULTAT DE CIÈNCIES

ECONÒMIQUES I EMPRESARIALS

Col·lecció d'Economia

Bootstrapping pairs in Distance-Based Regression[□]

Eva Boj del Val

Mª Mercè Claramunt Bielsa

Josep Fortiana Gregori

Adreça de correspondència:

Departament de Matemàtica Econòmica, Financera i Actuarial

Facultat de Ciències Econòmiques i Empresarials

Universitat de Barcelona

Avinguda Diagonal 690,

08034_Barcelona, ESPANYA

e-mails: *evaboj@ub.edu mmclaramunt@ub.edu fortiana@ub.edu*

[□] Treball financiat en part pel Projecte del Ministeri de Ciència i Tecnologia i el Fons FEDER MTM2005-02370.

Resum

La regressió basada en distàncies és un mètode de predicció que consisteix en dos passos: a partir de les distàncies entre observacions obtenim les variables latents, les quals passen a ser els regressors en un model lineal de mínims quadrats ordinaris. Les distàncies les calculem a partir dels predictors originals fent us d'una funció de dissimilaritats adequada. Donat que, en general, els regressors estan relacionats de manera no lineal amb la resposta, la seva selecció amb el test F usual no és possible. En aquest treball proposem una solució a aquest problema de selecció de predictors definint tests estadístics generalitzats i adaptant un mètode de bootstrap no paramètric per a l'estimació dels *p*-valors. Incluim un exemple numèric amb dades de l'assegurança d'automòbils.

Abstract

Distance-based regression is a prediction method consisting of two steps: from distances between observations we obtain latent variables which, in turn, are the regressors in an ordinary least squares linear model. Distances are computed from actually observed predictors by means of a suitable dissimilarity function. Being in general nonlinearly related with the response their selection by the usual F tests is unavailable. In this paper we propose a solution to this predictor selection problem, by defining generalized test statistics and adapting a non-parametric bootstrap method to estimate their *p*-values. We include a numerical example with automobile insurance data.

Key words: Distance-based regression; Predictors selection; Non-parametric bootstrap; Automobile insurance data.

JEL: C12, C14, C15, G22, C80.

1. Introduction

Distance-based regression (DBR) (Cuadras (1989), Cuadras and Arenas (1990), Cuadras *et al.* (1996)) is a prediction tool which can be applied directly to qualitative or mixed explanatory variables, while retaining compatibility with ordinary regression by least squares (LS), which appears as a particular case. Intuitively speaking, the model projects the vector of continuous responses onto a Euclidean space obtained by Metric Multidimensional Scaling (see, e.g., Borg and Groenen (1997)) from the observed predictors, which are nonlinearly mapped into a set of *latent*, i.e., non-observed, dimensions in this space. As predictors are nonlinearly related with the response, except for trivial or degenerated situations, they cannot be selected by the usual F tests.

The aim of this paper is to propose a new method for selecting predictors in the DBR model. To this end we define and study some properties of a significance test for predictors. Our constructed test statistic, *Q*, analogous to and a generalization of the classical F, appears through the concept of *geometric variability* (see Cuadras and Fortiana (2003)). Since, in general, the distribution of *Q* is unknown we estimate it via a non-parametric bootstrap technique: *bootstrapping pairs* (see Flachaire (1999)).

The organization of the paper is as follows: In Section 2 we outline the main characteristics of the DBR model, in Section 3 we define the new *Q* statistic, in Section 4 we adapt a bootstrapping pairs method to estimate its *p*-values, and in

Section 5 we illustrate the performance of the resulting predictor selection scheme by applying it to a real actuarial dataset.

2. The distance-based regression model

A continuous response Y is to be predicted from a set of p predictors, $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$, possibly a mixture of quantitative and qualitative variables. An n-vector, $\mathbf{y}$, contains the values of Y for an n-set Ω of individuals or cases. Let $\delta : \Omega \times \Omega \to \mathbb{R}^+$ be a distance function acting on the $\mathbf{w}_j$-coordinates (i.e., a function $\delta_{ij} = \delta\left(\mathbf{w}_i, \mathbf{w}_j\right)$ such that: $\delta_{ij} \geq 0$; $\delta_{ii} = 0$; $\delta_{ij} = \delta_{ji}$; $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$) and let $\mathit{\Delta} = \left(\delta_{ij}\right)$ be the related $n \times n$ matrix of inter-distances, the *predictor distance matrix.* $\mathit{\Delta}$ is called *Euclidean* if, for some integer *r,* we can find $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^r$, such that

$$\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \left(\mathbf{x}_i - \mathbf{x}_j\right) = \delta_{ij}^2, \quad 1 \leq i, j \leq n, \tag{2.1}$$

where the super-index T stands for matrix transposition. The $n \times r$ matrix X formed by stacking the n rows $\mathbf{x}_i^T$, the *Euclidean configuration* matrix, verifies that $G = HXX^T H$ is positive semi-definite (p.s.d.), where $H = I_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^T$ is the $n \times n$ centering matrix. Schoenberg’s theorem (see e.g. Thm. 14.2.1 in Mardia *et al.* (1979)) states that if

$$G = -\frac{1}{2} H \mathit{\Delta}^{(2)} H \tag{2.2}$$

is p.s.d., where $\mathit{\Delta}^{(2)}=\left(\delta_{ij}^2\right)$, then $\mathit{\Delta}$ is Euclidean, with $r=rank(G)\leq n-1$. In this case, any X such that $G=XX^T$ is a Euclidean configuration, automatically centered: $HX=X$. Note that in (2.2), the relation between G and $\mathit{\Delta}^{(2)}$ is $\mathit{\Delta}^{(2)}=\mathbf{g}^T\mathbf{1}_n^T+\mathbf{1}_n\mathbf{g}-2G$, where $\mathbf{g}$ is the row vector containing the main diagonal entries in *G*.

The DBR of $\mathbf{y}$ on $\mathit{\Delta}$ is defined as an LS regression with response $\mathbf{y}$ and matrix of predictors *X*, where *X* is a Euclidean configuration of $\mathit{\Delta}$. It can be proved that this definition is consistent, i.e., it does not depend on which Euclidean configuration *X* is chosen. Explicitly, the adjusted $\hat{\mathbf{y}}$ is given by:

$$\hat{\mathbf{y}}=X\hat{\boldsymbol{\beta}} \tag{2.3}$$

where $\boldsymbol{\beta}$ is such that $\left\|\mathbf{y}-\boldsymbol{\beta}X\right\|^2=\min$, i.e.,

$$\hat{\mathbf{y}}=X\left(X^TX\right)^{-1}X^T\mathbf{y} \tag{2.4}$$

where *X* is such that $\mathbf{1}_n^TX=\mathbf{0}$, $XX^T=G$, rank (*X*) = *r*, and we have assumed that **y** is centered.

Since $P=X\left(X^TX\right)^{-1}X^T$ is the unique orthogonal projector on the column space of *X*, which coincides with the column space of *G* (see, e.g., Rao (1973), 1.b.6, p. 27), $\hat{\mathbf{y}}$ does not depend on the choice of *X*. Also $P=G^+G=GG^+$ where G^+ is the Moore-Penrose g-inverse of *G*, hence

$$\hat{\mathbf{y}}=P\mathbf{y}\,. \tag{2.5}$$

Below we will need the following equality:

$$G^{+} = X\left(X^{T}X\right)^{-2}X^{T} \tag{2.6}$$

which can be proved by a direct computation.

As we mentioned above, a remarkable feature in DBR is that the ordinary LS model is recovered as a particular instance: Namely, when the explanatory variables actually belong to some $\mathbb{R}^{m}$ and we choose the natural Pythagorean, l^{2}, distance for δ. We refer the reader to Cuadras and Arenas (1990) and Cuadras *et al.* (1996) for a thorough discussion of the model and its properties. The prediction for a new individual $\{n+1\}$ is

$$\hat{\mathbf{y}}_{n+1} = \hat{\mathbf{x}}_{n+1}\hat{\boldsymbol{\beta}}, \tag{2.7}$$

where $\hat{\boldsymbol{\beta}}$ is as above and

$$\hat{\mathbf{x}}_{n+1} = \frac{1}{2}(\mathbf{g}-\mathbf{d})X\left(X^{T}X\right)^{-1} \tag{2.8}$$

is given by Gower's interpolation (Gower 1966, Gower and Hand 1996), from the row vector $\mathbf{d}$, which contains the n squared distances from $\{n+1\}$ to the previous ones. Taking (2.6) into account, we see that the resulting prediction

$$\hat{\mathbf{y}}_{n+1} = \frac{1}{2}(\mathbf{g}-\mathbf{d})X\left(X^{T}X\right)^{-2}X^{T}\mathbf{y} = \frac{1}{2}(\mathbf{g}-\mathbf{d})G^{+}\mathbf{y} \tag{2.9}$$

can be expressed in terms of distances only, that is, with no explicit dependence on *X*.

For the general case, if we have a mixture of quantitative, qualitative, and dichotomous variables, we can use the Euclidean distance based on Gower's similarity coefficient (Gower (1971)):

$$s_{ij} = \frac{\sum_{h=1}^{p_1}\left(1-\left|w_{ih}-w_{jh}\right|/G_h\right)+a+\alpha}{p_1+\left(p_2-d\right)+p_3} \qquad (2.10)$$

where p_1 is the number of continuous variables, *a* and *d* are the number of positive and negative matches, respectively, for the p_2 dichotomous variables, and α is the number of matches for the p_3 multi-state variables. G_h is the range of the *h*-th continuous variable. The square distance is $\delta_{ij}^2 = 1 - s_{ij}$ and $\Delta = \left(\delta_{ij}\right)$ is a Euclidean distance matrix (Gower and Legendre (1986)).

DBR allows a second Euclidean distance matrix Δ_{y} acting as the response. In our case,

$$\Delta_{\mathrm{y}}^{(2)} = \mathbf{g}_{\mathrm{y}}^T \mathbf{1}_n^T + \mathbf{1}_n \mathbf{g}_{\mathrm{y}} - 2G_{\mathrm{y}} \qquad (2.11)$$

where $G_{\mathrm{y}} = \mathbf{y}\mathbf{y}^T$. Linear prediction is given by

$$\hat{\Delta}_{\mathrm{y}}^{(2)} = \hat{\mathbf{g}}_{\mathrm{y}}^T \mathbf{1}_n^T + \mathbf{1}_n \hat{\mathbf{g}}_{\mathrm{y}} - 2\hat{G}_{\mathrm{y}}, \qquad (2.12)$$

where $\hat{G}_{\mathrm{y}} = \hat{\mathbf{y}}\hat{\mathbf{y}}^T$ is the projected inner product matrix, $\hat{G}_{\mathrm{y}} = PG_{\mathrm{y}}P$, and $\hat{\mathbf{g}}_{\mathrm{y}}$ contains its diagonal entries. This general formulation can be applied to predict qualitative or mixed responses (Fortiana and Cuadras (1998)) but here we adopt it just as a notational convenience.

3. Generalizing the F statistic: Geometric variability

In this section we define three quantities (3.2), (3.3) and (3.4), generalizing as many quantities usual in the study of ordinary LS regression and which, most importantly for our purposes, depend only on the inter-distances between individuals. We do this by making use of the concept of geometric variability, (3.1), defined in Cuadras and Fortiana (2003). Geometric variability is the extension of the concept of total variation in the field of distances.

1. The *geometric variability* of a distance matrix $\mathit{\Delta}$ is:

$$V(\mathit{\Delta}) = \frac{1}{2n^2}\mathbf{1}_n^T \mathit{\Delta}^{(2)} \mathbf{1}_n = \frac{1}{n} trG\,, \tag{3.1}$$

   where G is its associated inner product matrix. This quantity extends the concept of total variation, i.e., the trace of the covariance matrix.

2. The *coefficient of determination,* in terms of geometric variabilities,

$$R^2_{\mathrm{y,w}} = \frac{V(\hat{\mathit{\Delta}}_{\mathrm{y}})}{V(\mathit{\Delta}_{\mathrm{y}})}\,. \tag{3.2}$$

3. Given two models, $\hat{\mathbf{y}}^k = \mathrm{DBR}\left(\mathbf{y}, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\right)$ and $\hat{\mathbf{y}}^{k+1} = \mathrm{DBR}\left(\mathbf{y}, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k, \mathbf{w}_{k+1}\right)$, say, the *squared partial correlation coefficient*, in terms of geometric variabilities, is:

$$r^2_{\mathbf{y},\mathbf{w}_{k+1}|\mathbf{w}_1\mathbf{w}_2\cdots\mathbf{w}_k} = \frac{V^{k+1}(\hat{\mathit{\Delta}}_{\mathrm{y}}) - V^k(\hat{\mathit{\Delta}}_{\mathrm{y}})}{V(\mathit{\Delta}_{\mathrm{y}}) - V^k(\hat{\mathit{\Delta}}_{\mathrm{y}})}\,, \tag{3.3}$$

   where $V^k(\hat{\mathit{\Delta}}_{\mathrm{y}})$ comes from the first model and $V^{k+1}(\hat{\mathit{\Delta}}_{\mathrm{y}})$ from the second one.

4. Similarly, when comparing two models, the *test statistic*

$$Q \equiv Q(\mathbf{y},\mathbf{w}_{k+1} \,|\, \mathbf{w}_1\mathbf{w}_2\cdots\mathbf{w}_k) = \frac{V^{k+1}(\hat{\Delta}_{\mathrm{y}}) - V^{k}(\hat{\Delta}_{\mathrm{y}})}{V(\Delta_{\mathrm{y}}) - V^{k+1}(\hat{\Delta}_{\mathrm{y}})} \tag{3.4}$$

plays the role of the usual F test statistic to assess the significance of a new predictor, $\mathbf{w}_{k+1}$, added to a given set, $\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_k$. Indeed, since

$$RSS_0 = n\left[V(\Delta_{\mathrm{y}}) - V^{k}(\hat{\Delta}_{\mathrm{y}})\right] \text{ and } RSS_1 = n\left[V(\Delta_{\mathrm{y}}) - V^{k+1}(\hat{\Delta}_{\mathrm{y}})\right], \tag{3.5}$$

then $Q = \dfrac{RSS_0 - RSS_1}{RSS_1}$. When the explanatory variables actually belong to some $\mathbb{R}^m$ and δ is the natural l^2 metric, *Q* is proportional to F –degrees of freedom are not defined for a DBR model. For the statistic defined in (3.4) we have chosen a notation mimicking that of the partial correlation in (3.3), mainly because of similarity in their right hand sides.

4. Non-parametric Bootstrap

In order to test the null hypothesis that the addition of a new predictor, $\mathbf{w}_{k+1}$, to a given set, $\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_k$, does not significantly improve the current model:

$$H_0 : \mathbf{w}_{k+1} \text{ is not significant}, \tag{4.1}$$

we simulate the null distribution of *Q*, by adapting to the DBR context an appropriate version of non-parametric bootstrap. The basic principle is to generate *B* bootstrap samples by drawing with replacement from the observed dataset; for each of them the statistic of interest is calculated and percentiles can be evaluated from the B resulting values. For regression models, two possible paradigms are: *bootstrapping residuals,* in which each bootstrap sample of the response n-vector is derived from n resampled residuals, and *bootstrapping pairs* or *resampling cases,* in which each bootstrap sample consists of n response-predictor pairs from the original data (see Davidson and Hinkley (1997) for details, also Wehrens and van der Linden (1997)). The difference between the two methods is that in bootstrapping residuals the latent variables (or predictors in the DB model) are regarded as fixed. One assumes that the basic regression model is correct and that the residuals can be regarded as equal. If this is not the case for instance, when residuals have different variances or when errors are present in predictors bootstrapping residuals will yield erroneous results. The bootstrapping pairs paradigm, on the other hand, is less sensitive to wrong model assumptions. Furthermore, if the assumptions underlying bootstrapping residuals are met, bootstrapping pairs will yield approximately the same results. In this paper we concentrate on the bootstrapping pairs paradigm, adapting its data generating process (DGP) to the DB context.

The original form of the bootstrapping pairs DGP, proposed by Freedman (1981) was improved on by Flachaire (1999) with a resampling scheme that respects the null hypothesis of the test. Our adaptation of this refined version is as follows:

(a) Fit both models the one with k predictors, $\hat{\mathbf{y}}^k = \mathrm{DBR}\left(\mathbf{y}, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\right) = P^k \mathbf{y}$, and the one with k+1 predictors, $\hat{\mathbf{y}}^{k+1} = \mathrm{DBR}\left(\mathbf{y}, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k, \mathbf{w}_{k+1}\right) = P^{k+1}\mathbf{y}$ to the data. Then calculate Q $= Q(\mathbf{y}, \mathbf{w}_{k+1} \mid \mathbf{w}_1 \mathbf{w}_2 \cdots \mathbf{w}_k)$. Keep the two distance matrices generated, as they contain all the interdistances to be used in subsequent resamplings. In this way no further distance computations will be needed.

(b) Obtain the centered modified residuals under the alternative hypothesis H_1 by:

- computing the raw residuals under H_1, $\hat{\mathbf{r}} = \mathbf{y} - \hat{\mathbf{y}}^{k+1}$, and
- modifying and centering them by

$$\tilde{r}_i = \frac{\hat{r}_i}{\left(1 - p_i^{k+1}\right)^{1/2}} - \frac{1}{n}\sum_{s=1}^{n} \frac{\hat{r}_s}{\left(1 - p_s^{k+1}\right)^{1/2}} \quad \text{for } i = 1, \ldots, n, \tag{4.2}$$

where p_i^{k+1} is the (i,i)-th element of the main diagonal of the projector matrix P^{k+1}.

(c) Calculate the response under the null hypothesis for the bootstrap DGP, $\mathbf{y}_{H0}$, by adding the centered modified residuals to the null response:

$$\mathbf{y}_{H0} = \hat{\mathbf{y}}^k + \tilde{\mathbf{r}}. \tag{4.3}$$

(d) Randomly resample with replacement from the set $\left(\mathbf{y}_{H0}, W\right)$, putting probability 1/$n$ on each of the n observed data points, obtaining a bootstrap sample $\left(\mathbf{y}^*, W^*\right)$ of size n. Center $\mathbf{y}^*$.

(e) Fit both models, the one with *k* predictors and the one with *k*+1 predictors for the bootstrap sample just obtained, giving $\overset{\wedge}{\mathbf{y}}^{*k}$, $\overset{\wedge}{\mathbf{y}}^{*k+1}$. Calculate the bootstrap test statistic, $Q^{*} = Q(\mathbf{y}^{*}, \mathbf{w}_{k+1}^{*} \mid \mathbf{w}_{1}^{*}\mathbf{w}_{2}^{*} \cdots \mathbf{w}_{k}^{*})$.

(f) Repeat steps (a)—(e) *B* times. The relative frequency:

$$\frac{\#\left\{Q^{*} \geq Q\right\}}{B}. \tag{4.4}$$

is the bootstrap estimator of the *p*-value.

5. Numerical example: Automobile insurance data

In this section we illustrate the performance of the proposed DB predictor selection method. After describing the real dataset used in 5.1, we make two blocks of computations: First, in 5.2, in order to check the correct adaptation of the DGP to the DB context, we consider only continuous predictors and the natural l^2 distance, and compare and validate the results of resampling with those of the usual F test for LS regression. Second, in 5.3, we use the whole set of mixed predictors with the Gower similarity index (2.10) and make the complete selection process.

5.1. The dataset

Our application is in the *selection of tariff variables* in the rate-making process for automobile insurance. The response is the *expected claim amount* and the predictors are *observed risk factors*, i.e., quantities with a potential causal relationship with the response (Booth *et al.* (1999)). In insurance rate-making, the expected total claim amount per policyholder (*Pure premium*) is the product of the expected number of claims per policyholder by the expected claim amount, hence factors influencing each can be separately studied (Boj *et al.* (2005), Haberman and Renshaw (1996)).

In this paper, the empirical study is carried out using a portfolio from a Spanish automobile insurer, corresponding to compulsory civil liability insurance. We study factors which influence the claim amount (in ESP, with 1 EUR = 166.386 ESP) related to bodily injury. The actual data set consists of 455 claim amounts, belonging to the period 1/1/1996 - 1/1/1997, associated with the following eight risk factors:

Continuous:

Power = Power (in horse power) of the vehicle,

Vehicle age = Age of the vehicle on 1 January 1997,

Price = Original list price of the vehicle,

Age = Age of the main driver on 1 January 1997,

Categorical:

Sex = Sex of the main driver (2 levels),

Zone = Zone of use of the vehicle (10 levels),

Type = Vehicle type (4 levels),

Use = Vehicle usage (2 levels),

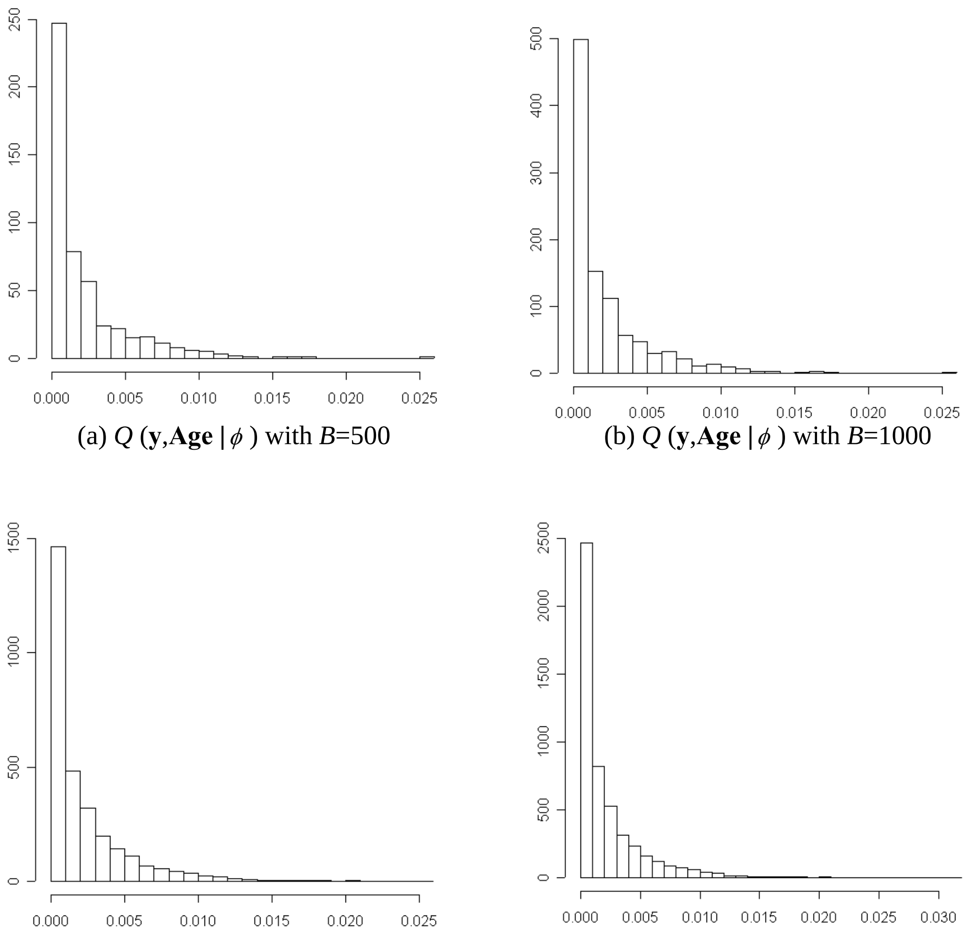
5.2. Bootstrap distribution of *Q*

In order to compare and validate the results of resampling with those of the usual F test for LS regression, we test the significance of entering each of the four variables, **Power**, **Vehicle age**, **Price**, **Age**, to the empty model, ϕ. In each case we compute the bootstrap distribution of the appropriate test statistic Q with B = 500, 1000, 3000 and 5000 resamples. In Table 1 we list the values of the statistics with the associated *p*-values. Columns 6 and 7 confirm that Q is indeed proportional to an F with 1 and 453 degrees of freedom (F = 453 Q).

Table 1. Q statistics (column 1) and associated *p*-values of entering the variables **Power**, **Vehicle age**, **Price**, and **Age** to the empty model ϕ, using the l^2 distance with B = 500, 1000, 3000 and 5000 resamples (columns 2 to 5) for the DBR model. And the F statistics and assymptotic *p*-values of the classical model (columns 6 and 7).

Candidate variable	Q	Estimated *p*-value B=500	Estimated *p*-value B=1000	Estimated *p*-value B=3000	Estimated *p*-value B=5000	F	*p*-value for F
Power	0.00010702	0.834	0.823	0.826	0.826	0.48480060	0.826
Vehicle age	0.00082011	0.540	0.546	0.551	0.546	0.37150983	0.542
Price	0.00002348	0.920	0.924	0.918	0.918	0.01063598	0.918
Age	0.00025328	0.750	0.734	0.742	0.737	0.11473584	0.735

Fig. 1 shows the histograms for one of these statistics, Q (**y**,**Age** $|\phi$), under four resampling sizes (the other three statistics behave similarly). In order to assess more precisely the similarity of its distribution to that of an F, in Table 2 we list areas under the right tail of the histogram for several Q values as compared with the corresponding F probabilities. The resulting information suggests that B = 1000 is an adequate sample size.



(a) Q (**y**,**Age** | ϕ) with B=500

(b) Q (**y**,**Age** | ϕ) with B=1000

(c) Q (**y**,**Age** | ϕ) with B=3000

(d) Q (**y**,**Age** | ϕ) with B=5000

Fig. 1. Histograms of bootstrap null distributions of the statistic Q(**y**,**Age** | ϕ) with B = 500, 1000, 3000 and 5000 resamples using the l^2 distance.

Table 2. Estimated areas of the right tail of the bootstrap null distribution of the statistic Q(**y**,**Age** | ϕ) with B = 500, 1000, 3000 and 5000 resamples (columns 2 to 5) using the l^2 distance, and right F probabilities of the classical model (column 7).

Q	Estimated area B=500	Estimated area B=1000	Estimated area B=3000	Estimated area B=5000	F	Classic probability
0.0005	0.628	0.623	0.641	0.637	0.2265	0.634
0.0050	0.142	0.134	0.132	0.129	2.2650	0.133
0.0070	0.080	0.072	0.073	0.073	3.1710	0.076
0.0090	0.042	0.040	0.041	0.041	4.0770	0.044
0.0095	0.032	0.031	0.034	0.034	4.3035	0.038
0.0125	0.010	0.010	0.011	0.012	5.6625	0.017
0.0200	0.002	0.002	0.002	0.002	9.0600	0.002

5.3. The predictor selection scheme

Now we illustrate the selection process taking into account the full set of mixed predictors and the distance derived from the Gower similarity index (2.10).

We perform a stepwise selection process as follows (see Tables 3 and 4): Starting with the null model with no predictors, the minimum *p*-value for adding one predictor corresponds to **Price** (Table 3, first column). **Power** is added to the resulting model in the next step (Table 3, second column). The corresponding tests for elimination are shown in Table 3, rows 1 and 2, where notations such as **Price|Power** stand for comparison of the model with **Price** and **Power** as predictors with the model with only **Power**. The test statistic for this comparison is $Q(\mathbf{y}, \mathbf{Price}|\mathbf{Power})$, as defined in Section 3. Successively, **Type** and **Use** are added in the same way (Table 3, columns 3 and 4) and the corresponding tests for deletion appear in Table 4, rows 3 and 4. Low significance of predictors is a known feature of bodily injury data, connected with the fact that claim amounts depend on risk factors that cannot be known *a priori*. An excessively strict observance of a conventional small significance level would lead us to consider no risk factors whatever. For a given portfolio under study it is better, as well as common practice, to accept the most significant predictor in each step, disregarding its numerical level.

Table 3. *p*-values for the four first inclusion phases of a stepwise predictor selection process taking into account the full set of mixed predictors, using the Gower similarity index and *B* = 1000 resamples.

	Variable already included			
Candidate variable	First step: ϕ	Second step: **Price**	Third step: **Price**, **Power**	Fourth step: **Price**, **Power**, **Type**
Power	0.474	**0.450**	--------	--------
Vehicle age	0.476	0.560	0.732	0.756
Price	**0.070**	--------	--------	--------
Age	0.196	0.790	1	1
Sex	0.446	0.656	0.700	0.664
Zone	0.792	0.544	0.732	0.730
Type	0.420	0.628	**0.625**	--------
Use	0.298	0.732	0.662	**0.630**
Added variable:	w(1) = **Price**	w(2) = **Power**	w(3) = **Type**	w(4) = **Use**

Table 4. *p*-values for the four first elimination phases of a stepwise predictor selection process taking into account the full set of mixed predictors, using the Gower similarity index and *B* = 1000 resamples.

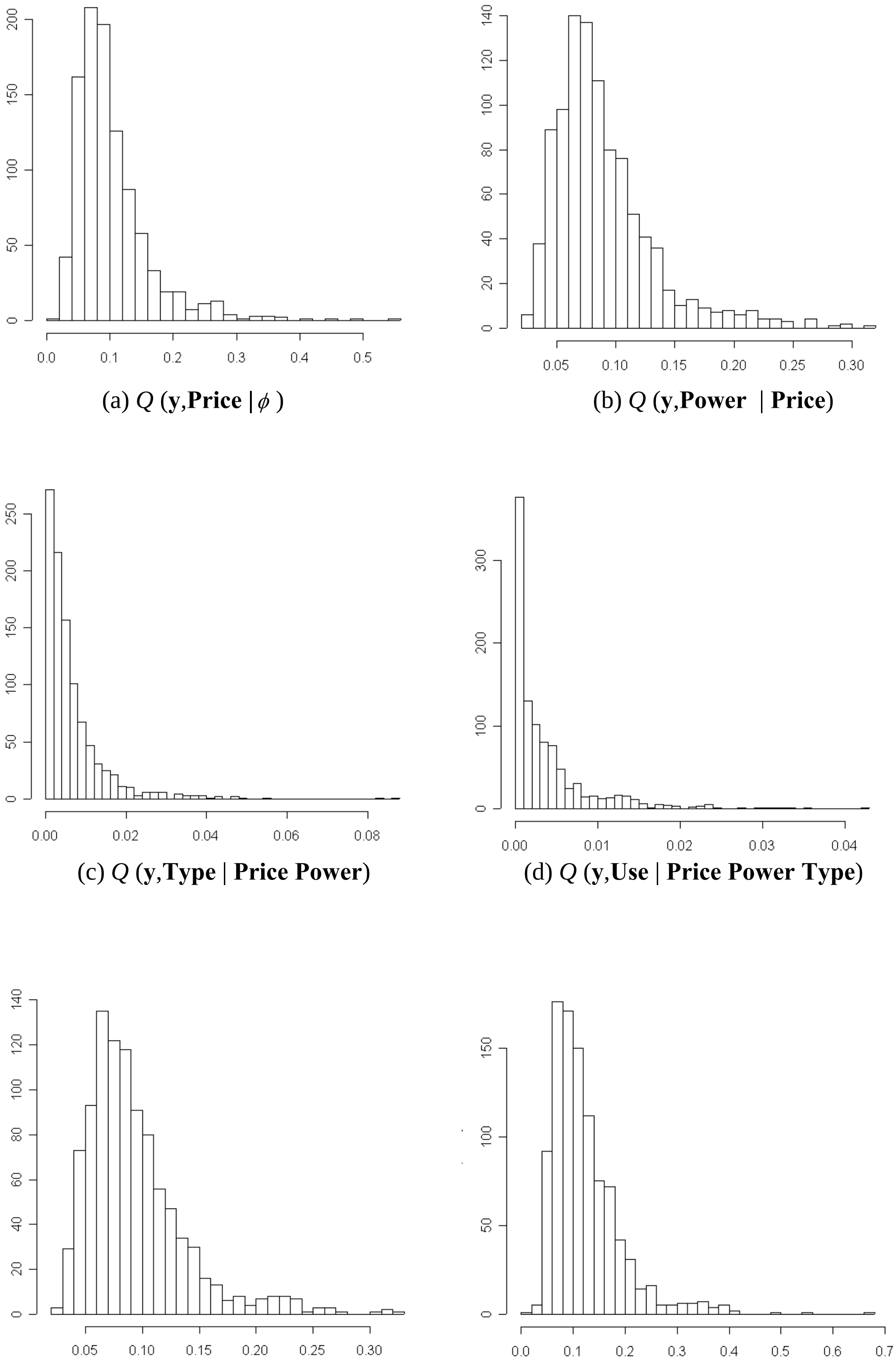
Excluded variable from a given set	*p*-value
First step: w(1) | ϕ	**0.070**

Second step:
w(2) | w(1)
w(1) | w(2)
	0.450
	0.120

Third step:
w(3) | w(1)w(2)
w(1) | w(2)w(3)
w(2) | w(1)w(3)
	0.625
	0.138
	0.484

Fourth step:
w(4) | w(1)w(2)w(3)
w(1) | w(2)w(3)w(4)
w(2) | w(1)w(3)w(4)
w(3) | w(1)w(2)w(4)
	0.630
	0.148
	0.504
	0.612

Hence the process suggests a model with **Price**, **Power**, **Type** and **Use**, the first four predictors appearing in the selection process. Its R^2, equal to 0.2306, is low, in agreement with the small predictive power of the known risk factors. In Fig. 2 we include some examples of estimated null distributions of *Q*. In all cases we use *B* = 1000 resamples.



(a) Q (**y**,**Price** | ϕ)

(b) Q (**y**,**Power** | **Price**)

(c) Q (**y**,**Type** | **Price Power**)

(d) Q (**y**,**Use** | **Price Power Type**)

(e) Q (**y**,**Power** | **Price Type**)

(f) Q (**y**,**Price** | **Power Type**)

Fig. 2. Histograms of bootstrap null distributions of the statistics Q(**y**,**Price** | ϕ), Q(**y**,**Power** | **Price**), Q(**y**,**Type** | **Price Power**), Q(**y**,**Use** | **Price Power Type**), Q(**y**,**Power** | **Price Type**) and Q(**y**,**Price** | **Power Type**) with B = 1000 resamples using the Gower similarity index.

While DBR cannot be regarded as a universally better replacement for classical prediction recipes with mixed explanatory variables, the results of the present paper provide it with a sound selection of variables tool and, as a consequence, render it a candidate alternative procedure.

For instance, a standard treatment for our dataset is a generalized linear model (GLM) with Gamma-distributed errors and logarithm link function where suitable dummies replace categorical predictors. In it the F test statistic based on deviances (see, e.g., Brockman and Wright (1992)) to include a first predictor in the model gives all *p*-values greater than 0.25. With our set of four predictors this GLM gives $R^2 = 0.0102$. Note, however, that for problems with a large number of categorical predictors the use of dummy indicators can eventually lead to sampling rarefaction, due to a curse of dimensionality effect, and to numerical unstabilities due to the singularity of the design matrix, whereas DBR is still appropriate when such problems arise.

6. Concluding remarks

In this paper we propose a method for selection of predictors in the DBR model. We construct a test statistic, *Q*, analogous to and a generalization of the classical F which appears through the concept of geometric variability. Since, in general, the distribution of *Q* is unknown, we estimate it via a non-parametric bootstrap technique, specifically by bootstrapping pairs. The two main contributions of the paper are: the definition of test statistic and the adaptation of bootstrapping pairs to the DBR context. Finally, we illustrate the performance of the resulting predictor selection scheme by applying it to a real actuarial dataset.

References

Boj, E., Claramunt, M. M., Fortiana, J. and A. Vegas (2005) “Bases de datos y estadísticas del seguro de automóviles en España: Influencia en el cálculo de primas”, *Estadística Española*, **160:47**, 539-566.

Booth, P., Chadburn, R., Cooper, D., Haberman, S. and D. James (1999) “*Modern Actuarial Theory and Practice*”, California: Chapman and Hall.

Borg, I. and P. Groenen (1997) “*Modern multidimensional scaling: theory and applications*”, New York: Springer.

Brockman, M. J. and T. S. Wright (1992) “Statistical Motor Rating: Making Effective Use of your Data”, *Journal of the Institute of Actuaries*, **119:3**, 457-543.

Cuadras, C. M. (1989) “Distance Analysis in discrimination and classification using both continuous and categorical variables”, in: *Statistical Data Analysis and Inference* (Y. Dodge ed.), Elsiever Science Publisher. North-Holland. Amsterdam, pp. 459-474.

Cuadras, C. M. and C. Arenas (1990) “A distance-based model for prediction with mixed data”, *Communications in Statistics: Theory and Methods*, **19**, 2261-2279.

Cuadras, C. M., Arenas, C. and J. Fortiana (1996) “Some computational aspects of a distance-based model for prediction”, *Communications in Statistics: Simulation and Computation*, **25:3**, 593-609.

Cuadras, C. M. and J. Fortiana (2003) “Distance based two multivariate sample tests”, *Preprint of the IMUB,* **334**.

Davidson, R. and D. Hinkley (1997) “*Bootstrap Methods and their Application*”, Cambridge, UK: Cambridge University Press.

Freedman, D. A. (1981) “Bootstrapping regression models”, *The Annals of Statistics*, **9**, 1218-1228.

Flachaire, E. (1999) “A better way to bootstrap pairs”, *Economics Letters*, **64**, 257-262.

Fortiana, J. and C. M. Cuadras (1998) “Generalized distance-based regression”, *Proceedings of the Classification & Psychometric Society Joint Meeting*, June 1998.

Gower, J. C. (1966) “Some distance properties of latent root and vector methods used in multivariate analysis”, *Biometrika*, **53**, 325-338.

Gower, J. C. (1971) “A general coefficient of similarity and some of its properties”, *Biometrics*, **27**, 857-874.

Gower, J. C. and P. Legendre (1986) “Metric and euclidean properties of dissimilarity coefficients”, *Journal of Classification*, **3**, 5-48.

Gower, J. C. and D. J. Hand (1996) “*Biplots*”, London: Chapman and Hall.

Haberman, S. and A. E. Renshaw (1996) “Generalized Linear Models and Actuarial Science”, *The Statistician,* **45:4**, 407-436.

Mardia, K. V., Kent, J. T. and J. M. Bibby (1979) “*Multivariate Analysis*”, London: Academic Press.

Rao, C. R. (1973) “*Linear statistical inference and its applications*”, New York: Wiley.

Wehrens, R. and van der Linden, W. E. (1997) “Bootstrapping principal component regression models”, *Journal of Chemometrics,* **11**, 157-171.