Application of bootstrap techniques in econometrics: the example of cost estimation in the automotive industry

Sandrine Juan, Frédéric Lantz
Renault, 1,avenue du Golf, F 78288 Guyancourt cedex – sandrine.juan@renault.com
IFP School, 228 avenue Napoléon Bonaparte, F 92852 Rueil-Malmaison cedex – frederic.lantz@ifp.fr

Abstract. The application of bootstrap methods to regression models helps approximate the distribution of the coefficients and the distribution of the prediction errors. In this paper, we are concerned with the application of the bootstrap techniques to determine prediction intervals on econometric models when the regressors are known. We investigate problems associated with its application: determination of the number of replications, choice of the method to calculate the least squares estimator (pseudo-inverse or inverse) and sorting algorithm of the statistic of interest. These investigations stemmed from the need to predict costs in the automotive industry in the earliest phases of the development of a new vehicle. Generally, the sample size is small and the error term of the model has not necessarily a Gaussian distribution. Consequently, the use of bootstrap techniques strongly improved prediction intervals by reflecting the original distribution of the data. Two examples (car engine and fuel tank) illustrate the application of such techniques.

Key words : bootstrap, pseudo-inverse, sorting methods, econometric forecast, car industry.

JEL Classification : C15, C53, L62

INTRODUCTION

Econometrics is extensively used for obtaining predictions in industrial applications today. The determination of confidence intervals of coefficients and prediction intervals depends on the assumptions inherent in the estimation methods and, in particular, assumptions on the distribution of the error term of the regression model. If these are no longer satisfied, standard prediction intervals are no longer usable.

The bootstrap proposed by Efron (1979) allows the approximation of an unknown distribution by an empirical distribution obtained by a resampling process. The application of bootstrap methods to regression models helps approximate the distribution of the coefficients (Freedman, 1981) and the distribution of the prediction errors when the regressors are data (Stine, 1985) or random variables (McCullough, 1996).

We are concerned here with the application of the bootstrap techniques to determine prediction intervals on econometric models when the regressors are known. These investigations stemmed from the need to predict costs in the automotive industry in the earliest phases of the development of a new vehicle. The general anticipated determination of the costs is part of the supply strategy of automotive manufacturers. As a rule, it aids the design of a new vehicle around a target price, and in particular, to make comparisons between various technical alternatives.

Among the different approaches available for making cost forecasts, the use of the econometric model is ideal in the early steps of an automotive project because it requires no detailed information. However, it raises difficulties connected with the small size of the data samples and the unknown distribution of the error terms of the regression models. In this context, bootstrap allows the use of an econometric approach for predictive purposes.

The first section briefly reviews the bootstrap principle on regression models. Section 2 goes on to address problems associated with its application: determination of the number of replications, choice of the method to calculate the least squares estimator (pseudo-inverse or inverse) and sorting algorithm of the statistic of interest. The next section deals with cost estimation in the preliminary project stage in the automotive industry and several applications of bootstrap (Section 3). Thus after estimating the cost of an engine, which only depends on a continuous variable, we present an econometric model of the cost of a fuel tank, involving a binary variable. The conclusion provides a summary of the results obtained and suggests a number of investigative channels.

1. BOOTSTRAP TECHNIQUES ON REGRESSION MODELS

Bootstrap is a resampling technique based on random sorts with retrieval in the data forming a sample. Used to approximate the unknown distribution of a statistic by its empirical distribution, bootstrap methods are employed to improve the accuracy of statistical estimations. Detailed presentations of this approach are proposed in particular by Hall (1992) and Efron and Tibshyrani (1993).

The use of bootstrap on regression models was initially broached by Freedman (1981). Jeong and Maddala (1993), Vinod (1993) and Veall (1998) offer syntheses of many developments and applications of bootstrap techniques in econometrics which subsequently appeared. Horowitz (1997) addressed the theoretical and numerical performance of bootstrap in econometrics. We shall briefly recall the principle of this resampling method and its application to regression models in Annex 1.

1.1 Bootstrap methods on regression models

The multiple linear regression model is denoted:

$$Y = X\beta + u \qquad (1)$$

where Y is a vector (*n*,1), X a matrix (*n*,*p*), β the vector of the coefficients to be estimated (*p*,1) and u the vector of random errors (*n*,1). A rank of observations i ($i = 1,\ldots,n$) of matrix X, corresponding to a line, is denoted X_i (1,*p*).

The parameter estimator β obtained by the ordinary least squares (OLS) method is expressed as $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the residuals as $\hat{u} = Y - X\hat{\beta}$.

For the rest of the discussion, we have adopted an approach in terms of bootstrap of residuals rather than an approach in terms of bootstrap by pairs, because we are not faced with the problem of heteroscedasticity (Flachaire, 1998).

The theoretical bootstrap model is as follows:

$$Y^* = X\hat{\beta} + u^* \tag{2}$$

where u^* is a random term obtained from the residuals $\hat{u}$ of the initial regression. At each iteration *b* (*b*=1,…,*B*), a sample $\{y_i^*\}_{i=1}^n$, of size (*n*,1), is created from the bootstrap model (2). Since the OLS residuals are smaller than the errors they estimate, the random term of the theoretical bootstrap model is constructed from the following transform residuals which have the same norm as the error terms u_i:

$$\tilde{u}_i = \frac{\hat{u}_i}{\sqrt{(1-h_i)}} - \frac{1}{n}\sum_{s=1}^{n}\frac{\hat{u}_s}{\sqrt{(1-h_s)}}$$

The theoretical bootstrap model is hence expressed as:

$$y_i^*(b) = X_i\hat{\beta} + \tilde{u}_i^*(b)\,, \quad \mathrm{i} = 1,\ldots,\mathrm{n} \tag{3}$$

where $\tilde{u}_i^*(b)$ is resampled from $\tilde{u}_i$.

Let us consider the random variable z_j, defined as $z_j = \dfrac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)}$, the standard confidence interval of β_j derives from the assumption according to which z_j is distributed according to a Student's distribution with *n-p* degrees of freedom. Thus for a confidence level $(1-2\alpha)$, this confidence interval takes the following form:

$$\left[\hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(1-\alpha),n-p}\ ,\ \hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(\alpha),n-p}\right] \tag{4}$$

where t are the percentile values (α) and $(1-\alpha)$ of the Student's distribution with *n-p* degrees of freedom.

The bootstrap confidence intervals are constructed from two percentile and percentile-t approaches. The first method, based exclusively on bootstrap estimations, is the simplest one for obtaining confidence intervals. For a level $(1-2\alpha)$, the percentile confidence interval for parameter β_j is given by:

$$\left[\hat{\beta}_j^*(\alpha B)\ ,\ \hat{\beta}_j^*((1-\alpha)B)\right] \tag{5}$$

where $\hat{\beta}_j^*(\alpha B)$ is the αB-th value (respectively $\hat{\beta}_j^*((1-\alpha)B)$ the $(1-\alpha)\,B$-th value) of the ordered list of the *B* bootstrap replications. The threshold values are hence selected so that

α% of the replications provide smaller (larger) $\hat{\beta}_j^*$ than the lower (upper) bound of the percentile confidence interval.

The percentile-t bootstrap procedure consists in estimating the distribution function of z_j directly from the data. This amounts to constructing a statistical table from the empirical distribution function of the B bootstrap replications z_j^*. This table is named the bootstrap table. The z_j^* are defined as:

$$z_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{s^*(\hat{\beta}_j^*)} \tag{6}$$

Let $\hat{F}_{z_j^*}$ be the empirical distribution function of z_j^*, the fractile containing α%, $\hat{F}_{z_j^*}^{-1}(\alpha)$, is estimated by the value $\hat{t}^{(\alpha)}$ such that $\#\left\{z_j^*(b) \leq \hat{t}^{(\alpha)}\right\} / B = \alpha$.
Finally, the percentile-t confidence interval for β_j is written:

$$\left[\hat{\beta}_j - s(\hat{\beta}_j) \cdot \hat{t}^{(1-\alpha)} \,,\, \hat{\beta}_j - s(\hat{\beta}_j) \cdot \hat{t}^{(\alpha)}\right] \tag{7}$$

Thus the percentile-t confidence interval is the bootstrap analogy of the standard confidence interval.

1.2 Bootstrap prediction intervals

After Stine (1985) and Breiman (1992), our working framework led us to use bootstrap to construct prediction intervals on regression models with fixed regressors, whose values are known (unconditional prediction). Note however that the construction of bootstrap prediction intervals on models with stochastic regressors is proposed by McCullough (1996).

For a new rank *f* of observations of explanatory variables X_f, the prediction of cost $\hat{y}_f$ is calculated from the regression model: $\hat{y}_f = X_f \hat{\beta}$.
Like the confidence intervals of the regression coefficients, the standard prediction interval derives from the assumption of Normality of the errors. Hence for a confidence level $(1-2\alpha)$, this standard prediction interval is written:

$$\left[\hat{y}_f - s_f \cdot t_{(1-\alpha),n-p} \,,\, \hat{y}_f - s_f \cdot t_{(\alpha),n-p}\right] \tag{8}$$

The use of bootstrap to determine the prediction intervals requires analysis of the prediction error distribution. Thus to preserve the same data generation process (DGP) for the estimations of the coefficients and predictions, the bootstrap prediction intervals are obtained with the bootstrap procedure for residuals. Similarly to the construction of the confidence intervals, two main methods are available for constructing the bootstrap prediction intervals: the percentile and percentile-t approach.

– Percentile prediction interval

The percentile method consists in using the bootstrap approximation of the prediction error distribution: $e_f = \hat{y}_f - y_f$, to construct a prediction interval of y_f. The theoretical bootstrap model corresponds to equation (3). The bootstrap replications of the future value y_f^*, for the new rank of observations X_f are generated by the same model (3) :

$$y_f^* = X_f \hat{\beta} + \tilde{u}_f^* \tag{9}$$

The error term $\tilde{u}_f^*$, like the $\tilde{u}^*$, is determined from a sort with retrieval in the empirical distribution of the transform residuals.

For each of the *B* bootstrap replications, we calculate the bootstrap estimator $\hat{\beta}^*(b)$, defined by equation (4). Thus the prediction and the bootstrap prediction error are written respectively:

$$\begin{gathered} \hat{y}_f^*(b) = X_f \hat{\beta}^*(b) \\ e_f^*(b) = \hat{y}_f^*(b) - y_f^*(b) \end{gathered} \tag{10}$$

Using equation (9), we can rewrite the bootstrap prediction error as:

$$e_f^* = \hat{y}_f^* - \hat{y}_f - \tilde{u}_f^* \tag{11}$$

Hence the latter inherently depends on the initial OLS prediction $\hat{y}_f$.

The *B* bootstrap replications of the prediction error give the empirical distribution of e_f^* : G^*. The percentiles of this empirical distribution, denoted $G^{*-1}(1-\alpha)$ and $G^{*-1}(\alpha)$, are then used to construct a bootstrap prediction interval.

A percentile prediction interval finally has the following form:

$$\left[\hat{y}_f - G^{*-1}(1-\alpha); \hat{y}_f - G^{*-1}(\alpha)\right] \tag{12}$$

– Percentile-t prediction interval

Similar to the confidence interval, the construction of the prediction interval, with the percentile-t method, implies the calculation, for each bootstrap sample, of the bootstrap estimator of the standard deviation. Thus to determine the percentile-t prediction intervals, the bootstrap estimator of the prediction standard deviation is necessary, for each of the replications. This is written:

$$s_f^* = s^* \cdot \sqrt{(1+h_f)} \tag{13}$$

where s^* is the bootstrap estimator of the standard deviation of the error terms and $h_f = X_f (X^T X)^{-1} X_f^T$.

The percentile-t procedure consists in constructing the statistics z_f^*, such that:

$$z_f^* = \frac{e_f^*}{s_f^*} = \frac{\hat{y}_f^* - \hat{y}_f - \tilde{u}_f^*}{s_f^*} \tag{14}$$

The bootstrap distribution of z_f^* defines the percentile-t bootstrap prediction interval. The percentiles $z_{f(1-\alpha)}^*$ et $z_{f(\alpha)}^*$, thus replace the critical values of the Student's distribution, considered in the standard prediction interval (cf. equation 8).
A percentile-t prediction interval is hence written:

$$\left[\hat{y}_f - s_f \cdot z_{f(1-\alpha)}^* ; \hat{y}_f - s_f \cdot z_{f(\alpha)}^*\right] \tag{15}$$

Note that, as for the confidence intervals of the coefficients, the percentile (1-α) of the distribution of z_f^* defines the lower bound of the prediction interval and vice versa for percentile (α).
A symmetrical distribution of z_f^* hence implies the symmetry of the percentile-t prediction interval. In the opposite case, however, the asymmetry is retranscribed in the inverse manner for the latter. For example, if z_f^* has a longer, right shifting distribution tail, the percentiles $z_{f(1-\alpha)}^*$ and $z_{f(\alpha)}^*$ are shifted towards the higher values of the bootstrap prediction errors, in comparison with the corresponding percentiles of a symmetrical distribution. Thus the resulting percentile-t prediction interval is shifted leftward, asymmetrical about the predicted OLS value. Hence its construction implies a sort of "automatic bias correction" and makes it possible, for a given confidence level, to accept lower predicted values than the symmetrical standard prediction interval.

2. APPLICATION OF BOOTSTRAP TO REGRESSION MODELS

The developments we propose for the use of the bootstrap technique are accordingly an extension of the work of Efron and Tibshirani (1993) and Booth and Sarkar (1998), Davidson and McKinnon (1988) for the determination of the number of bootstrap replications. Following McCullough and Vinod (1996), we addressed the choice of the method for calculating the OLS estimator by pseudo-inverse, as well as the sorting algorithm of the statistic of interest.

2.1 Number of bootstrap B replications

Efron and Tibshyrani (1993) recommended a number of bootstrap replications of around 25 to calculate the standard deviation of an estimator, and about a thousand for bootstrap confidence intervals. The construction of the bootstrap confidence interval (CI)[1] involved the fractiles (2.5%) and (97.5%) of the bootstrap empirical distribution of statistic z^*. We accordingly posed the following question: above what number of bootstrap replications do the estimations of fractiles of interest become invariant, according to the succession of random sorts performed? This amounts to analyzing the influence of *B* on these fractiles, to determine the number of replications required.

Intuitively, it seems logical that the estimation of the fractiles by a distribution demands a larger number of bootstrap samples than the calculation of the standard deviation of the estimator, for example. In fact, the latter depends on the tail of distribution, where few samples appear. The question boils down to determining the number of bootstrap replications above which the value of the fractiles can be considered stable.

The procedure employed consists, for a certain number of values of *B*, in performing a set of simulations (denoted *k*), aimed to judge the stability of the results, when the roots (starting values) of the random number generator are different. *k* = 100 simulations were made, which seems reasonable, in light of our investigations. The simulations were also made for the following numbers of bootstrap replications: *B* = 20, 30, 100, 500, 1,000, 5,000 and 10,000. Rather than analyzing the variability of each of the fractiles (2.5%) and (97.5%) of the distribution separately, we consider the range of the interval between these fractiles, which we name the "confidence interval of z^*". Thus our work relies on the analysis of the variability of the ranges of these CI, over 100 simulations, as a function of *B*.

In short, for each of the *k* simulations, $k = 1,\ldots,100$, the *B* bootstrap replications supply the empirical distribution of z^*, from which the fractiles of interest are extracted. For a given number of *B* replications, the simulations accordingly give 100 CI of z^* and their ranges. Their variability is then analyzed. The comparisons of the range distributions are made in pairs, for rising values of *B*. To do this, the characteristics of central values (mean, median) and dispersion are calculated and three criteria are examined: equality of the medians[2], equality of the variances, and thecoefficient of variation (CV). The first two criteria are the subject of statistical tests, respectively the Wilcoxon non-parametric test of equality of medians, and the Fisher tests of equality of variances of two independent samples. The variation in CV as a function of the number of replications is the subject of the analysis of the third criterion. The application and interpretation of this procedure is presented on the example of the engine cost estimation model.

[1] The process, presented for the construction of the bootstrap CI, applies identically to the construction of the bootstrap prediction interval.

[2] We have used the median as a central value indicator because, unlike the mean, it is unaffected by variations in the extreme values of the distribution.

2.2 Calculation of the estimator of the regression parameters

The calculation of the OLS parameters is usually performed (equation 2) through the inversion of the ($X^T X$) matrix. This could raise some numerical problems when the X matrix is ill-conditioned (Belsley, Kuh et Welsch, 1980). The calculation of the OLS estimator is therefore more accurate when it derives from the singular value decomposition of the X matrix :

$$\underset{(n,p)}{X} = \underset{(n,p)}{U} \; \underset{(p,p)}{D} \; \underset{(p,p)}{V^T} \qquad (17)$$

D is the singular values matrix of X. U is the orthogonal matrix of the p eigenvectors associated to the p non-zero eigenvalues of (XX^T) and V is the orthogonal matrix of the eigenvectors of ($X^T X$). Denoting $X^+ = V \; D^+ \; U^T$ the pseudo-inverse matrix of X, the OLS estimator is written as :

$$\hat{\beta} = X^+ y \qquad (18)$$

and its variance-covariance matrix is :

$$\hat{V}(\hat{\beta}) = s^2 \; V \; D^{-2} \; V^T$$

The projection matrix is given by :

$$I_n\text{-}X(X^T X)X^T = I_n - U \; U^T \qquad (19)$$

We have compared the two methods of calculation of the regression parameters (equations 2 and 18). After a first set of tests with Matlab (cf Juan, 1999), both methods have been directly programmed in Fortran. The most accurate algorithm for the inversion of the positive defined matrix ($X^T X$) is based on a cholesky decomposition (Seak, 1972). An analysis of the numerical results of this algorithm is given by Lantz (1983). We have used the algorithm of Golub et Reinsch (Forsythe and al., 1977) for the computation of the pseudo-inverse matrix X^+.

Multicollinearity has less consequences on the matrix conditionning when the data are scaled or centered and scaled (Belsley and al., 1980 – Belsley, 1984 – Erkel-Rousse, 1995). We have assessed these consequences for several dimensions of X (n=10,…,50, p=2,…,10). For each of them, we have generated 1000 matrices X such as X_1^T has a uniform distribution (0,1) and the following j vectors are built as $X_j^T = X_{j-1}^T + \nu$; ν has a uniform distribution (0, 0.001). We have measured the computation errors through the sum of the absolute values of the differences between XX^+ or $(X^T X)(X^T X)^{-1}$ and the identity matrix. Table 1 summarizes these errors of computation for several values of (n,p)[3]. This leads to choose the pseudo-inverse method with scaled data for the estimation of the OLS parameters. Nevertheless, a matrix inversion could be used with centered and scaled data when p is small.

[3] L’ensemble des résultats est disponible auprès des auteurs.

Table 1 – Computation error on the matrix product XX^+ and $(X^TX)(X^TX)^{-1}$

Dimension of X	XX^+ data scaled to 1	$(X^TX)(X^TX)^{-1}$ data scaled to 1	$(X^TX)(X^TX)^{-1}$ data centered and scaled
(50,2)	$<10^{-10}$	$<10^{-10}$	$<10^{-10}$
(50,4)	$<10^{-10}$	0.61×10^{-6}	0.11×10^{-6}
(50,8)	$<10^{-10}$	0.54×10^{-5}	0.59×10^{-6}
(50,10)	$<10^{-10}$	0.50×10^{-4}	0.50×10^{-5}

Subsequently, we have compared the computation time of residual bootstrap for these last two calculation methods (using the sorting algorithm described below). As previously, one thousand data samples have been generated for several values of (n,p). For each sample, the residual bootstrap technique associated to the percentile method have been applied with 1000 replications. Table 2 provides the computation time on a micro-computer (Pentium II, 300 Mhz). The running time is divided by approximately 2 when the pseudo-inverse calculation is performed.

Table 2 – Computation time for residual bootstrap associated to percentile method with 1000 replications.

Dimension of X	X^+ data scaled to 1	$(X^TX)^{-1}$ data centered and scaled
(10,2)	0.0103	0.0208
(50,4)	0.0518	0.0972
(90,6)	0.1001	0.1823

Unit : second

2.3 Sorting algorithm of the replications

The most two famous sorting algorithms are the sort by comparison of all the elements and the partition sort. This last method runs faster on large samples because it needs approximately $B\log_2 B$ comparisons to sort the B replications as the first one requires approximately $B^2/2$ operations.

The modified sorting algorithm that we propose is only looking for the determination of the α and 1-α percentiles. Thus, in the sort step, we are just working with $(\alpha B+1)$ bootstrap replications and not with all the (B) replications as in the classic algorithm. Two main steps can be distinguished.

First, we build a vector, denoted S_1 , which contains the first $(\alpha B+1)$ statistics $\hat{\theta}^*(b)$[4], sorted in a growing order. Therefore, $S_1(1)$ and $S_1(\alpha B+1)$ stand, respectively, for the smallest and the biggest values of the first $(\alpha B+1)$ bootstrap statistics. Then, S_1 is copied in an other vector, denoted S_2, with the same dimension.

[4] These are the value of the bootstrap statistics coming from the first b=1,…, $(\alpha B + 1)$ iterations.

Second, each (B–(αB+1)) remaining statistics is compared with the elements of S_1. The S_1 vector is used to look for the smallest elements of the B replications as follow. For each bootstrap iteration b , $b = \alpha B + 1,\ldots,B$, $\hat{\theta}^*(b)$ is compared with the largest value S_1. If it is greater or equal to this value, we go to the comparison with S_2. Otherwise, $\hat{\theta}^*(b)$ is compared to each element of S_1. If it lower or equal to the k-th element of S_1, S_1(k) is replaced by $\hat{\theta}^*(b)$ and for each index greater than k, the elements are moved forward the next upper one. Finally, the S_1 vector contains the (αB+1) smallest values of the B statistics. Subsequently, S_1(α B) is the α percentile of the $\hat{\theta}^*(b)$.

On a similar way, each (B–(αB+1)) remaining statistics is compared with the elements of S_2. The S_2 vector is used to look for the biggest elements of the B replications. Therefore, S_2 (1) stands for the $(1-\alpha)$ percentile of the bootstrap distribution. Note that we need to sort (αB+1) replications rather than (αB) to extract the $(1-\alpha)$ percentile because this is the first element of the S_2 vector of dimension (αB+1).

In this improved version of the sorting algorithm, we must distinguish the two steps in order to determine the number of comparisons. In the first step, the number of comparisons is equal to $\frac{\alpha B(\alpha B+1)}{2}$. In the second step, the number of operations has a lower bound when each statiscis is always lower than S_2(1) and is also always greater than S_1(αB+1) and has an upper bound when it is compared to the (αB+1) elements of S_1 and S_2. It is therefore contained between $[B-(\alpha B+1)]\times 2$ and $[B-(\alpha B+1)]\times(\alpha B+1)\times 2$.

The total number of comparisons of this algorithm is between $[(\alpha B+1)\times(\alpha B-1)]+B$ and $(\alpha B+1)\times(B-1)$; it is of order B^2 which is the same as for the classic algorithm. However, for B=5 000 and $\alpha = 0{,}025$, the number of operations is of order 12 millions with the classic algorithm and it is bouded by 20 000 and 600 000 with this improved version. This divides, at least, the number of operations by 20.

3. APPLICATIONS TO THE AUTOMOTIVE SECTOR

The automotive market is very competitive and the reduction of the costs is part of the supply strategy of automotive manufacturers. So, in the developped countries, the car markets are saturated and, nowadays there is a price competition between the car companies. F. Verboven (1996) analyzes the european car market in terms of oligopolistic competition. In the new industrialized countries where the sales are growing up, the purchasing power of the households is smaller and the automotive companies have to overcome their costs.

The cost estimation is carried out from the earliest phases of the development of a new vehicle to its production. In the first steps of a project (i.e. 36 months before manufacturing), it aids the design of a new vehicle around a target price, and in particular, to make comparisons between various technical alternatives (Juan, 1999). Thus, the cost prediction is an important target for the companies. Among the different approaches available for making cost forecasts, the use of the econometric model is ideal in the early steps of an automotive project because it requires no detailed information which is not available at this time.

However, it raises difficulties connected with the small size of the data samples and the unknown distribution of the error terms of the regression models. In this context, bootstrap allows the use of an econometric approach for predictive purposes by approximating this unknown distribution by the empirical one. We illustrate the use of bootstrap on such econometric model through two examples. Before, we present the costs that we study and the data that we use.

3.1 The data and general form of the models

The car consists of several thousand parts and accessories, forming sub-sets (or functions) like the engine, body, steering mechanism, etc. For a given vehicle model, different versions exist in terms of attachments, drive systems, etc. Thus to make cost forecasts for all the vehicles of the model concerned, cost estimation models are constructed for the sub-sets. The entity of which the cost is modeled hence usually corresponds to a sub-set of assembled parts. It is also important to clarify the perimeter of the costs analyzed. Our work addresses the production cost, called the manufacturing cost not including depreciation (PRF). This cost, which represents the expenditures by the company during the production phase of the good, is composed of purchases (of materials and outside produced parts) and a conversion value. The PRF accounts for about 60% of the total cost of a vehicle (which includes the costs of guarantee, logistics, etc). Note that the PRF not including depreciation does not include capital amortization.

The specificity of the cost models demands clarifying the industrial production context. Firstly, the costs of the different components of the database are standardized for an average manufacturing volume, representing the "usual" production conditions for the product family. We thereby discard effects of scale, by working on unit manufacturing costs. Secondly, the state of the art is considered as given. In fact, technological changes are chiefly reflected by new, more efficient machines. However, the cost parameter investigated does not include capital amortization. Thus the analytical framework leads us to specify cost models for a standardized production volume and a given state of the productive system. Moreover, our modeling needs lead us to use "cross-section" data of the cost of each of the components of the database at a given time. Hence the estimations are made in a static framework and do not include factors connected with technical progress.

Cost is thus explained by the pertinent technical characteristics of the product, at a given time of the technology, for purposes of cost prediction for new products. Our approach is therefore different from the one used in the application of bootstrap techniques on production boundary models (Simar, 1992). In the latter case, in fact, the problem is to assess the efficiency of a production unit with respect to an effective boundary, using bootstrap. Finally, note that while no "a priori" is set, concerning the form of the relation linking the cost with the descriptors, the linear or multiplicative forms are usually adopted.

3.2 Engine cost estimation model

The example covers the perimeter of the engine, with its electrical components. Roughly speaking, the engine is composed of the cylinder block, the mobile linkage (crankshaft, flywheel, etc), parts for the distribution (cam shaft, valves, etc) and the cylinder head (including its cover). The electrical equipment includes the starter, ignition coil, spark plugs, generator, etc.

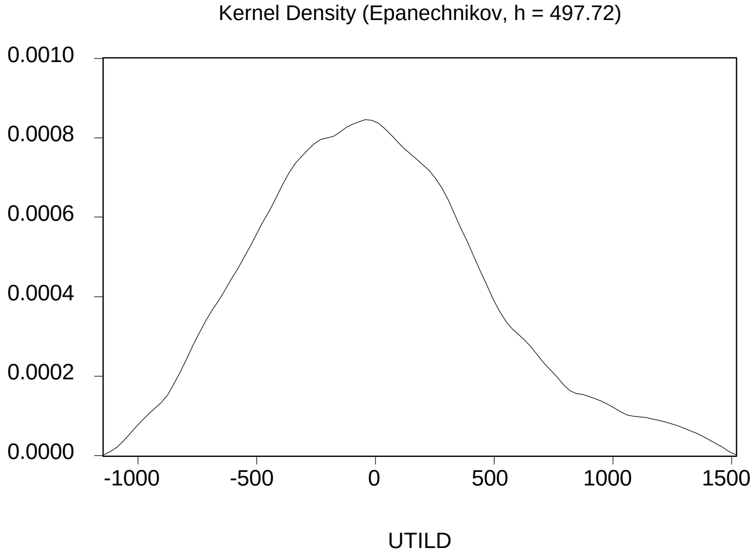
With the final assembly of the vehicle, the engine represents a function concerning which manufacture is usually an internal function of the company. Hence the cost data are PRF not including amortization. The engine also represents a non-negligible portion (about 20%) of the PRF of a vehicle.

A number of different engine families exist, chiefly distinguished by the carburation method (gasoline or diesel), the injection system (direct, monopoint, etc) and the volumetric displacement. The study is concerned with engines with a displacement larger than 1,700 cm^3, forming a uniform working sample of 15 engines. In the technical perimeter of the engine, the presence of additional components (support, pipes, etc) are imposed by accessories like power steering and air conditioning. These may or may not be present on the engine, depending on the duty rendered by the vehicle on which they are to be mounted.

In the econometric equation, the cost is a function of the volumetric displacement of the engine. Thus, the regression includes an explanatory variable and an intercept. The model is estimated by the ordinary least squares (OLS) method. The value of R^2 is 0.97 and the estimated coefficients[5] of the regression are significantly different from zero, for a first type risk of 5%. From the White test F(2,12)=0.66, we do not reject homoscedasticity.

The transform residuals of the regression $\tilde{u}$, from which the samples in the residuals bootstrap procedure are formed, are shown in Figure 1. The tail of distribution is longer towards the right, reflecting the presence of high positive residuals. In fact, the residual corresponding to a 1,783 cm^3 engine, which we denote $\hat{u}_6$, the "extreme" residual, has the highest value of all the residuals. This engine, as well as the second in the sample, are equipped with air conditioning and correspond to the highest residuals of the regression.

Figure 1 Density core estimator of transform residuals



[5] For reasons of confidentiality, the values of the estimated parameters are not given.

The procedure for determining the number of bootstrap replications needed is then applied (cf. section 2.1). The results presented concern the example of the displacement coefficient $\boldsymbol{z}_1^*$.

– Number of B replications

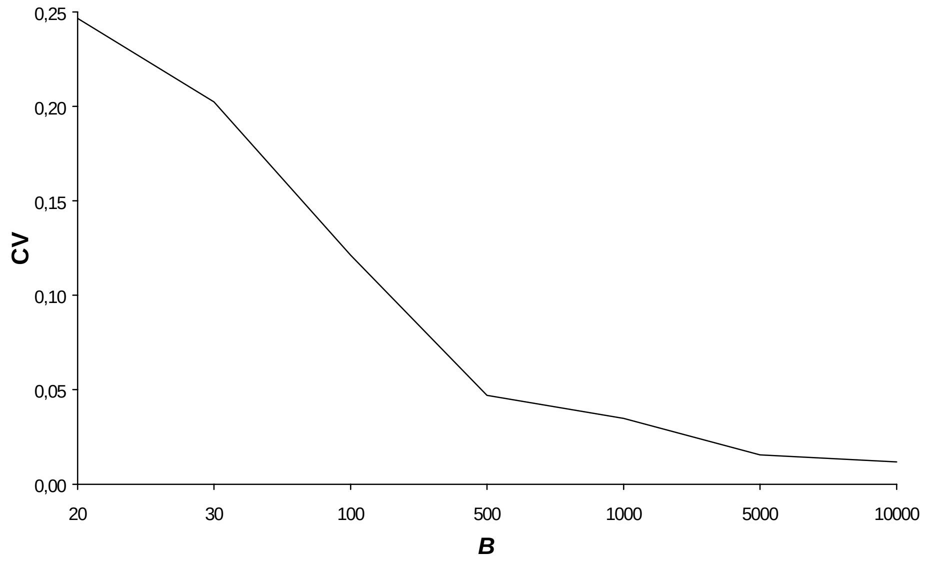
Table 3 Tests of CI ranges of $\boldsymbol{z}_1^*$

B	**Median of ranges**	**Wilcoxon test**	**P-marg.**	**Standard deviation of ranges**	**Fisher test**	**P-marg.**
20	3.910			0.973		
30	4.307	3.676	0.000	0.889	1.197	0.372
100	4.224	0.864	0.388	0.520	2.924	0.000
500	4.273	1.130	0.258	0.203	6.565	0.000
1 000	4.304	0.458	0.647	0.150	1.830	0.003
5 000	4.301	0.030	0.976	0.067	5.019	0.000
10 000	4.317	1.604	0.109	0.051	1.727	0.007

The results of the Wilcoxon tests for the equality of the medians and the Fisher tests for the equality of the variances of the ranges of the confidence intervals of $\boldsymbol{z}_1^*$ do not positively establish a value of *B* from which we can accept the assumption of equality of the medians and the variances of the CI (cf table 3).

The change in percentage standard deviation as a function of the number of replications is shown in Figure 2. The CV declines sharply for low values of *B*, and then slackens progressively. Thus from *B* = 5,000, it is then stabilized. This third criterion finally enables us to select the number of bootstrap replications *B* = 5,000, from which the CI ranges of z_1^* (and hence the fractiles (2.5%) and (97.5%) of the distribution) can be considered stable. Similarly, *B* = 5 000 replications are selected to construct the bootstrap prediction intervals.

Figure 2 Percentage standard deviation of CI ranges of $\boldsymbol{z}_1^*$



– Cost prediction

The fitting of the regression is now used to predict the PRF of a new engine, with displacement of 1,900 cm^3.

Table 4 OLS predictions and standard and bootstrap prediction intervals

Engine displacement in cm^3	**OLS prediction**	**Standard prediction interval**			
		2.5%	**97.5%**	**Range**	**Form**[6]
1900	9133.42	8160.67	10106.20	1945.53	1.00

Engine Displacement in cm^3	**Percentile prediction interval**				**Percentile-t prediction interval**			
	2.5%	**97.5%**	**Range**	**Form**	**2.5%**	**97.5%**	**Range**	**Form**
1900	8419.66	10179.57	1759.91	1.47	8370.00	10271.97	1901.96	1.49

Units: French francs

For the bootstrap prediction intervals, we show two construction methods: percentile and percentile-t. The percentile intervals possess a much smaller range than the percentile-t or standard intervals. Thus the percentile method yields too "optimistic" (too small) intervals and does not appear relevant for this type of investigation. In fact, the prediction error is not a pivot statistic and the bootstrap inference may prove wrong in this case. Note that the bootstrap prediction intervals are shifted towards the higher cost values (form higher than 1) compared with the standard intervals.

– Marking the "extreme" residual in bootstrap replications for prediction

To explain the asymmetry of the bootstrap prediction intervals, we analyze the impact of the sorting of the extreme residual $\hat{u}_6$ in the bootstrap DGP, on the bootstrap prediction error distribution. To do this, we examine, for each of the replications, the residual $\tilde{u}_f^*$ of the theoretical bootstrap model of the prediction (cf. equation 13) and check whether or not it corresponds to the residual $\hat{u}_6$. Table 3 shows the means and standard deviations of z_f^* for these two alternatives (the bootstrap statistic of the normed prediction error), for the five thousand replications.

[6] $form = \frac{upp - \hat{y}_f}{\hat{y}_f - low}$, where upp and low correspond effectively to the lower and upper bounds of the boostrap prediction interval, where $\hat{y}_f$ is the OLS prediciton.

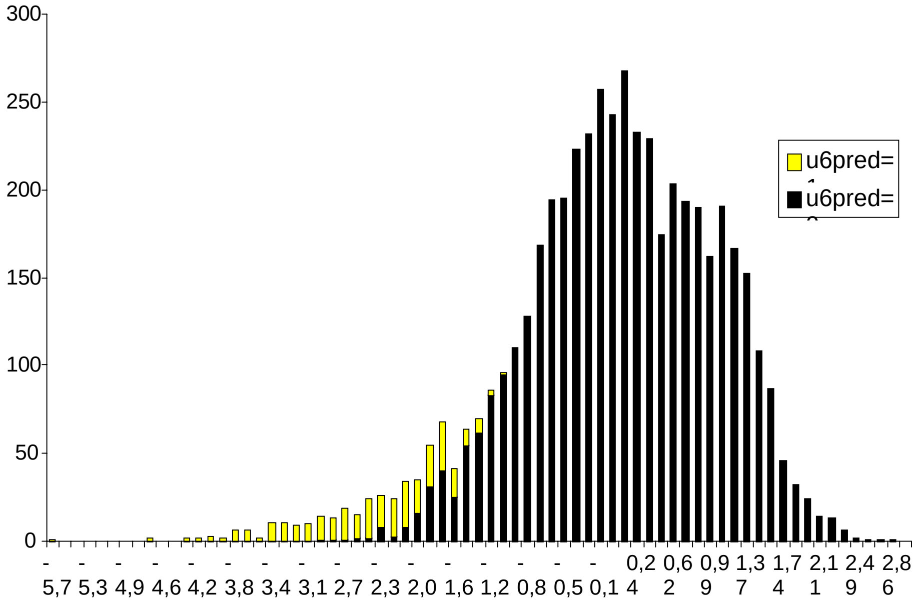
(*forme* = form; *sup* = upper; *inf* = lower)

Table 5 Characteristics of $\boldsymbol{z^*_f}$

Engine	$\boldsymbol{z^*_f}$	$\boldsymbol{\tilde{u}^*_f \neq \hat{u}_6}$	$\boldsymbol{\tilde{u}^*_f = \hat{u}_6}$	**Total**
1900 cm^3	Mean	0.144	-2.472	-0.022
	Standard deviation	0.896	0.745	1.092
	Batch quantity	4682	318	5000

For the cost prediction of a 1,900 cm³ engine, Figure 3 shows the distribution of the bootstrap prediction error, distinguishing the case in which the extreme residual is sorted.

Figure 3 Distribution of statistic $\boldsymbol{z^*_f}$, for a 1,900 cm^3 engine



The bootstrap distribution of the normed prediction error z^*_f appears to be strongly asymmetrical towards the negative values. Moreover, we observe (Figure 3) that they asymmetry is due to the sorting of the extreme residual $\hat{u}_6$, in the theoretical bootstrap prediction model. In fact, since $e^*_f = \hat{y}^*_f - \hat{y}_f - \tilde{u}^*_f$, if $\tilde{u}^*_f = \hat{u}_6$ (high positive value), the bootstrap prediction error is strongly negative, and this is found it the histogram. Thus the resulting prediction interval is shifted towards the higher values of cost, since the fractile (2.5%) of the distribution determines the upper bound of the interval, and vice versa for the fractile (97.5%). Note that this process is repeated for the distribution of the statistic z^*_f of other predictions (cf Juan, 1999).

3.3 Fuel tank cost estimation model

The technical perimeter of the product investigated is the fuel tank equipped with the fuel gauge and the suction pump. The tank is made of plastic, by a blowing process. It is part of a larger subassembly of the vehicle: the fuel circuit.

The equipped tank is an example of the function of the vehicle, of which the design and manufacture are completely externalized and the domain of equipment suppliers. Hence the tank is entirely an "outside produced part". The cost of the tank, to the automobile manufacturer, is the purchase price. This includes the supplier's profit margin. Yet it can be considered constant, by reference to terms of sale corresponding to constant volumes, whether for the automotive manufacturer or for the suppliers, whose products make up our sample.

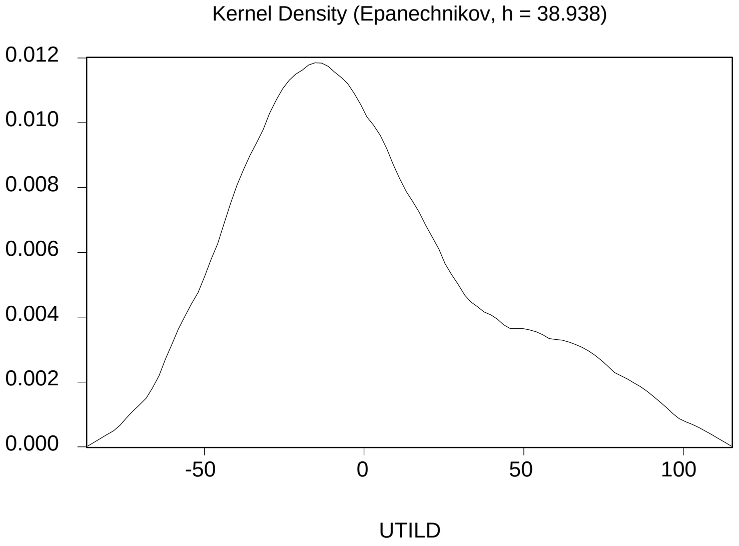
It is also important to distinguish diesel tanks from gasoline tanks. The discrimination is technically explained by the presence of a built-in pump in the suction system, on gasoline tanks, which is not mounted on diesel tanks. Gasoline is also much more volatile than diesel, and pollution control standards impose a permeability limit on the tank. These desiderata explain the specific anti-evaporation treatment. Note that the cost of the fluorination process primarily depends on the depollution standard, and marginally on the capacity of the tank. Thus gasoline tanks display an extra cost compared with their diesel counterparts. This extra cost is also variable. In fact, the architecture and design criteria may entail a double well (one for gauging and one for suction) depending on the vehicle.

In the econometric model, the fuel tank cost is a function of a capacity variable and a dummy variable which is zero if the tank contains diesel and one if it contains gasoline. The regression is estimated by OLS. The value of R^2 is 0.95 and the estimated parameters of the capacity and fuel variables are significant, for a first type risk of 5. From the White test F(3,11)=1.29, we do not reject homoscedasticity.

We also performed a Fisher test, to assess the validity of an overall model, compared with two distinct regressions, on each type of tank (gasoline and diesel). This is equivalent to testing the equality constraint of the capacity coefficient, for the two sub-samples F(1,11)=0.54 which is lower than the critical value at 5%. Thus the test does not allow us to discard the zero hypothesis of a single regression for gasoline and diesel tanks.

Figure 4 shows the kernel density estimator of the transform residuals $\tilde{u}$, without any distinction of the type of tank. The tail of distribution is much longer towards the right, reflecting the presence of high positive residuals. In fact, the residuals corresponding to observations Nos. 2 and 8 (gasoline tanks) display high values. After analysis, it turns out that contrary to the other gasoline tanks, these are equipped with a double well, imposed by the architectural requirements of the vehicle. These residuals are called "extreme" residuals.

Figure 4 Kernel density estimator of transform residuals



– Bootstrap procedures for classic and stratified residuals

After having presented the regression model and analyzed the residuals, it is necessary to examine the process of construction of the bootstrap samples, which, in the presence of a dummy variable in the regression, displays certain specificities. In fact, the bootstrap procedure for "classic" residuals, as presented in section 1.2, does not respect the structure of the initial sample. In the bootstrap DGP, the residuals (gasoline or diesel) are randomly assigned to the capacity of a diesel tank, for example. Intuitively, this procedure does not seem satisfactory since the approach underlying the bootstrap consists in generating artificial samples as similar as possible to the initial sample. This means that in the presence of one (or more) explanatory variables of the qualitative type in the model, we propose an adapted version, called the "bootstrap procedure for stratified residuals".

The construction of the bootstrap sample is discussed in detail below.

Let $\tilde{u}$ be the dimension vector *n* = 15, the transform residuals of the regression. This vector is split into two sub-samples: $\tilde{u}_1$ of size $n_1 = 8$, composed of residuals associated with gasoline tanks, and $\tilde{u}_2$, of size $n_2 = 7$, of the 7 residuals associated with the diesel tanks. We perform respectively 8 (7) random sorts with retrieval in $\tilde{u}_1$ ($\tilde{u}_2$) to create $\tilde{u}_1^*$ ($\tilde{u}_2^*$). The concatenation of the two vectors $\tilde{u}_1^*$ and $\tilde{u}_2^*$ then forms the vector of the bootstrap residuals resampled "by strata": $\tilde{u}^*$ with dimension *n* = 15. The rest of the procedure is identical to the bootstrap procedure for standard residuals, with the bootstrap residuals vector $\tilde{u}^*$ of dimension *n* = 15. The two residual (classic and stratified) bootstrap procedures are applied and compared during the construction of bootstrap prediction intervals.

– Cost prediction

The cost predictions and their intervals are calculated for diesel and gasoline tanks, with capacity of 60 liters (Tables 6, 7 and 8).

Table 6 OLS predictions and standard prediction intervals

Tank	OLS prediction	Standard prediction interval			
		2.5%	97.5%	Range	Form
60 l Diesel	260.24	182.74	337.74	157.00	1.00
60 l Gasoline	507.50	430.37	584.63	154.26	1.00

Units: French francs

Table 7 Bootstrap prediction intervals of classic residuals

Tank	Percentile prediction interval				Percentile-t prediction interval			
	2.5%	97.5%	Range	Form	2.5%	97.5%	Range	Form
60 l Diesel	203.20	342.22	139.02	1.44	203.33	354.27	150.94	1.65
60 l Gasoline	449.90	588.30	138.40	1.40	451.73	599.26	147.53	1.64

Units: French francs

Table 8 Bootstrap prediction intervals of stratified residuals

Tank	Percentile prediction interval				Percentile-t prediction interval			
	2.5%	97.5%	Range	Form	2.5%	97.5%	Range	Form
60 l Diesel	220.52	308.91	88.39	1.22	215.33	312.98	97.65	1.17
60 l Gasoline	441.50	594.94	153.44	1.32	446.19	619.38	173.19	1.82

Units: French francs

The prediction intervals obtained with the bootstrap procedure for classic residuals are similar, in terms of range, to the standard intervals. Their forms are asymmetrical towards the higher costs, in the same way for gasoline and diesel tanks. Hence this procedure, which randomly reassigns the residuals, retranscribes the asymmetry of their distribution equally on the two types of tank. In fact, this asymmetry, caused by the gasoline tanks, should not be transferred to the diesel tanks. The bootstrap procedure for stratified residuals supplies prediction intervals with smaller ranges (about 100 FF) for diesel tanks and wider ranges (175 FF) for gasoline tanks. These intervals are also symmetrical for diesel tanks and strongly asymmetrical towards the right for gasoline tanks.

– Marking of "extreme" residuals in bootstrap replications for prediction

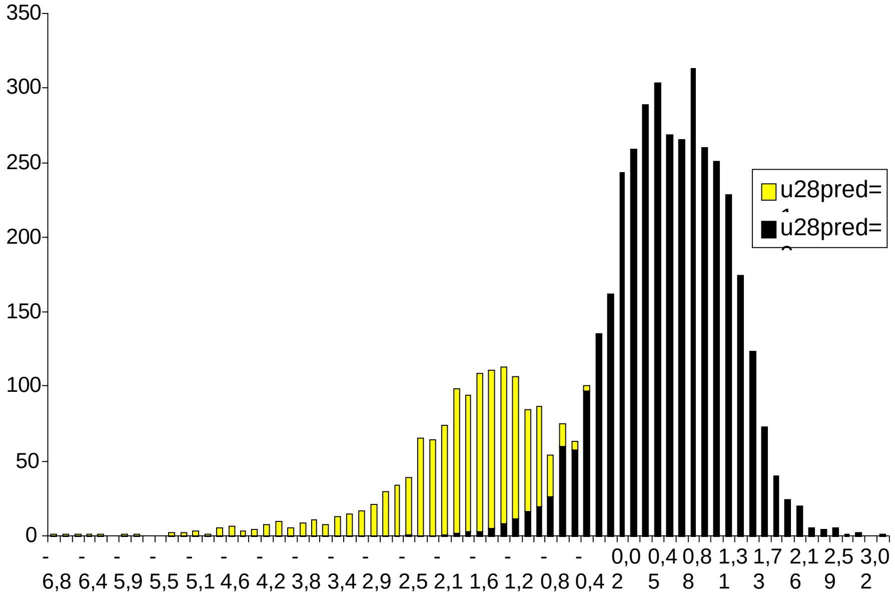
The impact of the extreme residual sort on the distribution of the bootstrap prediction error is analyzed, by checking whether for each replication, the residual $\tilde{u}_f^*$ of the theoretical bootstrap model of the prediction corresponds to $\hat{u}_2$ or $\hat{u}_8$. Table 9 shows, as a function of $\tilde{u}_f^*$, the means and standard deviations of the bootstrap statistic of the normed prediction error, for five thousand bootstrap replications.

Table 9 Characteristics of $\boldsymbol{z}_f^*$ for a 60 liter tank

	60 liter tank	$\boldsymbol{z}_f^*$	$\tilde{\boldsymbol{u}}_f^* \neq (\hat{\boldsymbol{u}}_2$ ***ou*** $\hat{\boldsymbol{u}}_8)$	$\tilde{\boldsymbol{u}}_f^* = (\hat{\boldsymbol{u}}_2$ ***ou*** $\hat{\boldsymbol{u}}_8)$	**Total**
Bootstrap of classic residuals	Diesel	Mean	0.256	-1.984	-0.028
		Standard deviation	0.793	0.761	1.086
		Batch quantity	4365	635	5000
	Gasoline	Mean	0.265	-2.062	-0.033
		Standard deviation	0.790	0.895	1.119
		Batch quantity	4360	640	5000
Bootstrap of stratified residuals	Diesel	Mean	0.010	0.002	0.008
		Standard deviation	0.733	0.772	0.743
		Batch quantity	3723	1277	5000
	Gasoline	Mean	0.603	-1.974	-0.035
		Standard deviation	0.685	0.897	1.337
		Batch quantity	3763	1237	5000

Figure 5 shows the empirical bootstrap distribution of the prediction error for a 60 liter gasoline tank, when the bootstrap procedure for residuals is stratified. The bootstrap distribution of the normed prediction error z_f^* appears to be highly asymmetrical towards the negative values. This asymmetry is due to the sorting of extreme residuals $\hat{u}_2$ or $\hat{u}_8$, in the theoretical bootstrap prediction model (cf. Figure 5). In fact, since $e_f^* = \hat{y}_f^* - \hat{y}_f - \tilde{u}_f^*$, if $\tilde{u}_f^* = \hat{u}_2$ ou $\hat{u}_8$ (high positive values), the bootstrap prediction error is strongly negative. This process is reproduced for the distribution of the statistic z_f^* of other predictions (cf Juan, 1999).

Figure 5 Distribution of statistic z_f^*, for a 60 liter gasoline tank (bootstrap of stratified residuals)



CONCLUSION

The use of bootstrap on regression models allows to approximate the unknown distribution of the error of prediction by its empirical distribution. With a small size of the data samples and an unknown distribution of the error terms of the regression models, the bootstrap techniques are very useful for an econometric approach.

The determination of the number of replications is based on the coefficient of variation of the width of the condidence interval. The use of a pseudo-inverse matrix rather a traditional matrix inverse gives more accurate estimations of the regression parameters when the data are ill-conditioned. Moreover, this method runs faster as it does not require data centering. A modified sorting algorithm allows to quickly sort the statistics obtained from the bootstrap replications. Only distribution tails are sorted and each element of the series is compared to the retained percentiles.

The first example developed illustrates the use of bootstrap techniques on the simplified model of a regression with a single explanatory variable, when the size of the sample is small. The analysis of the bootstrap prediction intervals shows that bootstrap serves to retranscribe the asymmetry of the distribution of the residuals in the prediction intervals. In fact, the latter are shifted towards the higher costs in comparison with the standard intervals, thereby allowing higher values for the cost prediction of a new engine. In this context, the use of bootstrap techniques accordingly allows a better retranscription of the information contained in the initial sample for the cost prediction intervals of a new product.

The second example served to describe an application of adjusted bootstrap techniques, in the case of modeling in the presence of a dummy variable. This method helps to construct symmetrical prediction intervals for diesel tanks and asymmetrical towards the higher cost values for gasoline tanks. This results from the asymmetry, both in the distribution of the prediction error and that of the extra cost connected with the gasoline tank. Thus the use of the bootstrap procedure of stratified residuals, when the information is available, makes it possible to take account, in the confidence and prediction intervals, of potential additional costs for gasoline tanks, imposed by the vehicle's architectural criteria.

Further developments on the use of bootstrap techniques are investigated. They concern the estimation of possible non-linear cost functions which have to be distinghuised from asymmetric relashionships.

References

Belsley D., Kuh E., Welsch R., 1980, “ *Regression diagnostics, identifying influential data and sources of collinearity* ”, ed. John Wiley and Sons, New-York

Belsley D., 1984, “ Demeaning conditioning diagnostic through centering ”, *The American Statistician*, Vol. 38, n°2, p 73-93

Breiman L. , 1992, “The little bootstrap and other methods for dimensionality selection in regression : X-fixed prediction error ”, *Journal of the American Statistical Association*, Vol. 87, p. 738-754.

Booth J.G., Sarkar S., 1998, “Monte-Carlo approximation of bootstrap variances”, *The American Statistician*, Vol. 52, n°4, p 354-357

Davidson R., McKinnon J., 1998, “Bootstrap tests : how many bootstraps ? ”, Document de travail, Université de la Méditerranée, GREQAM

Efron B. , 1979, “ Bootstrap methods : another look at the jacknife ”, *Annals of Statistics*, Vol. 7, p. 1-26.

Efron B., Tibshyrani R.J. , 1993, “ *An introduction to the bootstrap* ”, ed. Chapman and Hall, New-York

Erkel-Rousse H., 1995 ,“ Détection de la multicolinéarité dans un modèle linéaire ordinaire : quelques éléments pour un usage averti des indicateurs de Belsley, Kuh et Welsch ”, *Revue de statistique appliquée*, Vol. 18, n°4, p 19-42

Flachaire E., 1998, “ les méthodes du bootstrap et l’inférence robuste à l’hétéroscédasticité“, Thèse de doctorat, Université de la Méditerranée, GREQAM

Forsythe G.E., Malcolm M.A., Moler C.B., 1977, “ *Computer methods for mathematical computations* ”, Prentice-Hall

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

Hall P., 1992, “ *The bootstrap and edgeworth expansion* ”, ed. Springer Verlag, New-York

Horowitz J.L., 1997, “ Bootstrap methods in econometrics : theory and numerical performance”, “ Advances in economics and econometrics : theory and application ”, Vol. 3, Cambridge University Press, p.188-222.

Jeong J., Maddala G.S., 1993, “ A perspective on application of bootstrap methods in econometrics ”, *Handbook of Statistics*, Vol. 11, Amsterdam : North-Holland, p. 573-610.

Juan S., 1999, “ Les modélisation économétriques d’estimation de coût dans l’industrie automobile: l’apport des techniques de bootstrap“, Thèse de doctorat, ENSPM-Université de Bourgogne

Lantz F., 1983, “ *Mise en œuvre de la régression linéaire : calcul et stabilité de la matrice (X’X) inverse*”, Les cahiers du 3^{e} cycle économétrie , Université de Paris X, p. 34-47

McCullough B.D., 1996, “ Estimating forecast intervals when the exogenous variable is stochastic ”, *Journal of Forecasting*, Vol. 15, p. 293-304.

McCullough B.D., Vinod H., 1993, “ Implementing the single bootstrap : some computational considerations”, *Computational Economics*, Vol. 6, p. 1-15.

Seaks T., 1972, “ Computer algorithms - syminv : an algorithm for the inversion of a positive definite matrix by the Cholesky decomposition”, *Econometrica*, Vol. 40, n°5

Simar L., 1992, “ Estimating efficiencies from frontier models with panel data : a comparison of parametric, non parametric and semi-parametric methods with bootstrapping ”, *The Journal of Productivity Analysis*, Vol.3, p. 171-203.

Stine R.A ., 1985, “ Bootstrap prediction intervals for regression ”, *Journal of The American Statistical Association*, Vol. 80, p. 1026-1031.

Veall M.,R., 1998, “Applications of the bootstrap”, *Handbook of Applied Economic Statistics*, Vol. 155, ed. Aman U., Giles D.E.A.

Verboven F., 1996, “ International price discrimination in the european car market”, *Rand Journal of Economics*, Vol. 27, p 240-268

Vinod H., 1993, “Bootstrap methods : applications in econometrics”, *Handbook of Statistics*, Vol. 11, North-Holland, p.629-661.

Annex 1 - Bootstrap principle and application to the regression model

A.1 Principle

According to the bootstrap principle, by repeating the resampling several times in the initial data, one can construct the empirical bootstrap distribution function of a statistic of interest. This satisfactorily approximates the true distribution of the statistic, which is itself unknown. The process of constructing the empirical bootstrap distribution function of an estimator is described in detail below.

Consider a sample i.i.d. $\{y_i\}_{i=1}^{n}$, of a random variable *y* of unknown distribution *F*. Let us try to determine the distribution $\hat{\theta}(F)$ of an estimator $\hat{\theta}$ of a parameter θ of *F*. The aim is hence to construct a statistical table[7] of approximate values of $\hat{\theta}(F)$. To do this, we have *n* observations of the sample, hence of the empirical distribution function $\hat{F}_n$.
In theory, it is possible to construct a table of $\hat{\theta}(\hat{F}_n)$: *T*, conditional at $\{y_i\}_{i=1}^{n}$, by calculating $\hat{\theta}$ on each of the *n*-samples taken with retrieval in $\{y_i\}_{i=1}^{n}$. However, n^n such samples exist and this procedure can only be implemented if *n* is very small. In practice, we shall take *B* samples ($b = 1,\ldots,B$), named bootstrap samples, to construct a table extracted from *T*. On each of the bootstrap samples, we calculate the value $\hat{\theta}^*(\mathrm{b})$ of the statistic $\hat{\theta}$. The bootstrap table of $\hat{\theta}^*$ is hence a sub-table *T*, conditional on the data sample. This concept of dependence on data is important for understanding the bootstrap process. The bootstrap table only in fact applies for the initial sample. For a new sample, it is therefore necessary to construct a new bootstrap table, specific to this sample.

According to the theory of Glivenko-Cantelli (G-C), for a sample of random variables i.i.d. of unknown distribution *F*, if the size of the sample *n* tends towards infinity, the empirical distribution function of the sample uniformly converges almost definitely towards distribution *F*. Thus the bootstrap table of $\hat{\theta}^*$, conditional to $\{y_i\}_{i=1}^{n}$, supplies a good approximation of the distribution $\hat{\theta}(F)$.

A.2 Bootstrap methods on regression models

The multiple linear regression model is denoted:

$$Y = X\beta + u \tag{A-1}$$

[7] « Statistical table » means an empirical version (on B bootstrap replications) of the sampling distribution of $\hat{\theta}$.

where Y is a vector (n,1), X a matrix (n,p), β the vector of the coefficients to be estimated (p,1) and u the vector of random errors (n,1). A rank of observations i $(i = 1,\dots,n)$ of the matrix X, corresponding to a line, is denoted X_i (1,p). The parameters estimated by the ordinary least squares (OLS) method $\hat{\beta}$ and the residuals $\hat{u}$ are defined as: $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{u} = Y - X\hat{\beta}$.

– Bootstrap of residuals

The theoretical bootstrap model is as follows:

$$Y^* = X\hat{\beta} + u^* \qquad \text{(A-2)}$$

where $\hat{\beta}$ is the OLS estimator and u^* is a random term taken from the residuals $\hat{u}$ of the initial regression, whereof we describe the construction below.

The application of the bootstrap procedure consists in repeating the following steps *B* times:

1) At each iteration *b* (*b*=1,…,*B*), a sample $\left\{y_i^*\right\}_{i=1}^{n}$, of size ($n$,1), is created from the bootstrap model (A-2). We accordingly have a new pair (Y^*, X) on which we can make an estimation of the regression parameters. Since the OLS residuals are smaller than the errors they estimate, a transform is necessary to establish the random term of the theoretical bootstrap model. Hence following Freedman (1981), this is constructed with the following transform residuals ($\hat{u}_i$ is divided by a factor proportional to the root of its variance) which are of the same norm as the error terms u_i:

$$\tilde{u}_i = \frac{\hat{u}_i}{\sqrt{(1-h_i)}} - \frac{1}{n}\sum_{s=1}^{n}\frac{\hat{u}_s}{\sqrt{(1-h_s)}}$$

The theoretical bootstrap model is hence the following:

$$y_i^*(b) = X_i\hat{\beta} + \tilde{u}_i^*(b)\,, \quad i = 1,\dots,n \qquad \text{(A-3)}$$

where $\tilde{u}_i^*(b)$ is resampled from $\tilde{u}_i$. The new dependent bootstrapped variables $y_i^*(b)$ are hence constructed from calculated values $\hat{y}_i$ and from $\tilde{u}_i^*(b)$: $y_i^*(b) = \hat{y}_i + \tilde{u}_i^*(b)$.

1) The OLS estimation procedure is applied to the regression model (A-3) to obtain the bootstrap estimator. For the *b*-th sample, this is written:

$$\hat{\beta}^*(b) = (X^T X)^{-1} X^T Y^*(b) \qquad \text{(A-4)}$$

Steps 1) and 2) are repeated *B* times (*b*=1,…,*B*). The empirical bootstrap distribution functions of $\hat{\beta}^*$ and the statistics obtained from $\hat{\beta}^*$ are then constructed.

In applying a bootstrap procedure on residuals, the explanatory variables *X* are considered as fixed. Hence this procedure is valid if the *X* are really fixed and if the errors satisfy the classic assumptions of OLS[8]. In consequence, it is not correct if the latter are heteroscedastic. By applying the residuals bootstrap procedure to construct the theoretical bootstrap model, different error terms are associated with different explanatory variables. Thus the bootstrap residual resampling procedure is unable to respect this relation. In such cases, the theoretical bootstrap model is constructed by a different procedure, called "bootstrap by pairs".

– Bootstrap by pairs

This second bootstrap approach to regression models consists in directly resampling in the original data, from pairs (y_i, X_i). Note however that the simultaneous resorting of (y_i, X_i) introduces a correlation between the regressors and the errors of the bootstrap data generation process. Hence bootstrap by pairs, in this simple form[9], does not satisfy the hypothesis of exogeneity of the regressors in the bootstrap DGP.

The application of the bootstrap by pairs procedure consists in repeating the following steps *B* times:

1) At each iteration *b* (*b*=1,…,*B),* the vector Y^* and the matrix of explanatory variables X^* are constructed, by making *n* random sorts with retrieval[10] of pairs (y_i, X_i), in the original sample. Hence if the error term u_i associated with X_i has a large variance, the relation is preserved in the bootstrap sample.

2) An OLS estimation of the coefficients of the bootstrap regression model is then performed:

$$\hat{\beta}^*(b) = (X^{*T}(b)X^*(b))^{-1}X^{*T}(b)Y^*(b) \qquad \text{(A-5)}$$

Unlike the residuals bootstrap procedure, note that the matrix of explanatory variables $X^*(b)$ is different at each iteration *b*.

The *B* replications $\hat{\beta}^*$ accordingly supply the empirical bootstrap distribution function.

Thus the *B* independent replications bootstrap, obtained by the bootstrap procedures discussed above, supply a random sample of $\hat{\beta}^*$ which is used to estimate the bootstrap distribution of $\hat{\beta}$. This allows the construction of the bootstrap confidence intervals and the parameters of the regression model.

[8] The errors have a zero mathematical likelihood, are homoscedastic and non-autocorrelated.

[9] The *wild bootstrap* is an appropriate method in the presence of heteroscedacticity, which satisfies the hypothesis of exogeneity of regressors.

[10] Note that Friedman (1981) considers bootstrap samples of size *m* different from *n*.

A.3 Construction of bootstrap confidence intervals

Let us briefly recall the form of the standard confidence intervals of the regression coefficients, which depend on the hypothesis of Normality of the error terms, before presenting the bootstrap confidence intervals, which depend exclusively on the data. The developments are made for the element j of the vector of parameters: β.

Consider the random variable z_j defined as $z_j = \dfrac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)}$, the standard confidence interval of β_j derives from the assumption according to which z_j is distributed by a Student's distribution with *n-p* degrees of freedom. Thus for a confidence level $(1-2\alpha)$, the standard confidence interval assumes the form:

$$\left[\hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(1-\alpha),n-p}\ ,\ \hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(\alpha),n-p}\right] \qquad \text{(A-6)}$$

where t are the values of the percentiles (α) and $(1-\alpha)$ of the Student's distribution with *n-p* degrees of freedom.

The bootstrap confidence intervals are now presented, through their two main methods of construction: the percentile and percentile-t approaches.

– Percentile method

This method, based exclusively on the bootstrap estimations, is the simplest for obtaining confidence intervals. The first step is to sort the estimations $\hat{\beta}_j^*(b)$ by increasing order, for the *b*=1,…,*B* replications. Let $\hat{\beta}_j^*(1)$ be the smallest and $\hat{\beta}_j^*(B)$ the largest.
For a level $(1-2\alpha)$, the percentile confidence interval for parameter β_j is accordingly given by:

$$\left[\hat{\beta}_j^*(\alpha B)\ ,\ \hat{\beta}_j^*((1-\alpha)B)\right] \qquad \text{(A-7)}$$

$\hat{\beta}_j^*(\alpha B)$ represents the αB-th value (respectively $\hat{\beta}_j^*((1-\alpha)B)$ the $(1-\alpha)$ *B*-th value) of the ordered list of *B* bootstrap replications. The threshold values are hence selected so that α% of the replications have supplied $\hat{\beta}_j^*$ smaller (larger) than the lower (upper) bound of the percentile confidence interval.

– Percentile-t method

This second approach for constructing the confidence interval follows a procedure very similar to the one used for the standard confidence interval of β_j. The percentile-t bootstrap procedure consists in estimating the distribution function of z_j directly from the data. This amounts to constructing a statistical table from the empirical distribution function of the B bootstrap replications z_j^*. This table is named the bootstrap table. The z_j^* are defined as:

$$z_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{s^*(\hat{\beta}_j^*)} \tag{A-8}$$

In comparison with the percentile method, an additional calculation is necessary in this approach. In fact, for each of the bootstrap replications, it is necessary to calculate the estimated bootstrap standard deviation $s^*(\hat{\beta}_j^*)$.

Let $\hat{F}_{z_j^*}$ be the empirical distribution function of z_j^*, the fractile at α%, $\hat{F}_{z_j^*}^{-1}(\alpha)$, is estimated by the value $\hat{t}^{(\alpha)}$ such that $\#\left\{z_j^*(b) \leq \hat{t}^{(\alpha)}\right\} / B = \alpha$.

Finally; the percentile-t confidence interval for β_j is written:

$$\left[\hat{\beta}_j - s(\hat{\beta}_j)\cdot \hat{t}^{(1-\alpha)}\, ,\, \hat{\beta}_j - s(\hat{\beta}_j)\cdot \hat{t}^{(\alpha)}\right] \tag{A-9}$$

Thus the percentile-t confidence interval is the bootstrap analogue of the standard confidence interval.

In short, the percentile-t confidence interval substitutes the threshold values of the bootstrap table for the critical values of the Student's distribution used in the standard interval. Note that these threshold values may be very different. This difference is commensurate with the deviation of the error distribution (unknown) from the Normal distribution. Moreover, we find that the values of the percentiles (α) and ($1-\alpha$) of the Student distribution, which are inherently symmetrical, directly implies the symmetry of the standard confidence interval about the estimation $\hat{\beta}_j$. By contrast, the values $\hat{t}^{(\alpha)}$ and $\hat{t}^{(1-\alpha)}$ of the bootstrap table may be asymmetrical and accordingly lead to asymmetrical confidence intervals about $\hat{\beta}_j$. This consideration of a possible asymmetry is a major advantage of the bootstrap confidence intervals.