BOOTSTRAP SIMULATION MODELS IN ENVIRONMENTAL ECONOMICS

ALBEANU, Grigore
Faculty of Mathematics and Informatics
Spiru Haret University
g.albeanu.mi@spiruharet.ro

BURTSCHY, Bernard
ENST, INFRES, France
burtschy@enst.fr

POPENȚIU-VLĂDICESCU, Florin
UNESCO IT Chair
University of Oradea
popentiu@imm.dtu.dk

GHICA, Manuela
Faculty of Mathematics and Informatics
Spiru Haret University
m.ghica.mi@spiruharet.ro

Abstract
This paper describes the usage of some bootstrap simulation models for modelling loss distributions and for general dynamic financial analysis applied in environmental economics. Two cases studies, coming from forest economics, are presented.

Keywords: *Boostrap, LDA, DFA, loss distribution, regression, portfolio management*

AMS classification: 62P20

1. Introduction

Computer-intensive methods for estimation assessment provide valuable information concerning the adequacy of applied probabilistic models. The bootstrap method is an extensive computational approach for understanding empirical data and is based on resampling and statistical estimation. Bootstrap is a simple but powerful Monte Carlo method to assess statistical accuracy or to estimate a distribution from data samples. The name may come from the phrase "Pull yourself up by your bootstrap" having the following meaning 'Improve your situation by your own efforts' [26]. The methods are suitable for any level of modelling being useful for fully parametric, semi-parametric, and completely nonparametric analysis. These approaches are not only in use by statisticians, but also are applied anywhere statistics can be used: life sciences, business, social sciences, econometrics, reliability etc. Moreover, bootstrap is a powerful tool, especially when only a small data set is used to predict the behaviour of systems or processes. For the aim of this paper we outline

the basic bootstrap principle, the application of bootstrap sampling for accuracy estimation and the method of simultaneous confidence bands for uncertainty management.

Let X be a random variable and F the cumulative distribution function of the variable X. The Bootstrap method, proposed by Efron [10, 11, 13], is useful, at least, for the estimation of: the distribution function of a random variable R(X, F); a functional relation V(F), or the accuracy of a statistics s obtained from a sample $(X_1, X_2, ..., X_n)$ of size n ($\geq$1) from X. For this investigation, the accuracy, describes the variability of s when independent estimations s(1), s(2), ..., of the statistics s, are obtained by resampling.

The bootstrap technique uses the sample (X_1, X_2, …, X_n) to obtain the sampling cumulative distribution function $F_n(x)$ in order to replace the true cumulative distribution function F: $F_n(x)$ = 1/n cardinal $\{x_i \leq x;\ 1 \leq i \leq n\}$. To repeatedly simulate bootstrap samples $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$ from F_n, random number generators should be used according to the Monte Carlo approaches [12], [17]. Then, for each bootstrap sample, it is recalculated: the distribution function of the random variable $R(X^*, F_n)$; the functional relation $V(F_n)$ or $V(F_n^*)$; and the statistics s*(). The accuracy of the statistics s can be derived under an appropriate statistical inference study on the sequence s*().

The bootstrap resampling can be realized in various ways. Uniform resampling and the importance resampling are the mostly used [1], [2], [12], [20], [22]. Uniform resampling assumes that the measurement (observed) values are uniformly sampled from some process, while the importance resampling algorithm assumes the generation of sampling values according to a probability distribution $\{(x_i, p_i)$: I = 1., 2, …, n} such as every p_i is a nonnegative real number, and $p_1+p_2+\ldots+p_n=1$.

The general framework for obtaining simultaneous confidence bands for non-linear models (estimated from data) using resampling techniques is presented in [1, 2]. Let g_+, g_- be two known nonnegative models (real functions which may depend on the input data, denoted by D), and u_+, u_- be nonnegative scale factors. The confidence band for the model h (used to explain a relation such as $y_i = h(x_i, \theta) + \varepsilon_i$, I = 1, 2, …, n, θ as estimated parameter; or h is a non-parametric estimated model) is the region:

$$R=\left\{(x,y): x\in D, h\left(x,\hat{\theta}\right)-\hat{V}u_{-}g_{-}(x)\leq y\leq h\left(x,\hat{\theta}\right)+\hat{V}u_{+}g_{+}(x)\right\}, \qquad (1)$$

where $(x_i, y_i)_{i\geq 1}$ are observations, θ is a parameter vector, $\hat{\theta}$ is a least squares estimate of θ, $(\varepsilon_i)_{i\geq 1}$ are random variables distributed independently and identically with an unknown cumulative distribution function (cdf) F, but with mean zero and unknown variance σ^2, and $\hat{V}$ is an estimate of σ.

Obtaining the region R asks for solving the following problem: given the models g_+, g_-, and $\alpha\in(0,1)$, search for the non-negative real numbers u_-, u_+ such that

$$P\left(h\left(x,\hat{\theta}\right)-\hat{V}u_{-}g_{-}(x)\leq h(x,\theta)\leq h\left(x,\hat{\theta}\right)+\hat{V}u_{+}g_{+}(x), x\in D\right)=1-\alpha. \qquad (2)$$

There exist various alternatives for choosing g_+ and g_- [1] [2]. A common approach considers the models $g_+(x) = g_-(x) = 2$ for all $x \in D$. Two approaches related to the scale factors u_+, u_i are important enough: $u_+ = u_-$ (symmetrically simultaneous confidence bands); or the sum $u_+ + u_-$ must have a minimal value.

2. **Operational Risk**

Measuring the operational risk of an enterprise or financial institution requires the Loss Distribution Approach (LDA), which makes use of the exact operational loss frequency and severity distributions [16], [19], [24]. Financial aided decision uses not only LDA, but also the Dynamic Financial Analysis (DFA).

Mainly, the LDA has three essential components: a distribution of the annual number of losses (the frequency), a distribution of the monetary losses (the severity), and an aggregate loss distribution that combines the two. The first step in applying LDA method is dedicated to the understanding of the structure and characteristics of the data. According to [25], a careful Exploratory Data Analysis (EDA) is necessary to be performed before modelling the shape of the data by statistical tools.

The most used approach in building loss distribution models consists in fitting the parametric distributions to the loss severity and the frequency data [16], [19]. For modelling, there are available both simple parametric distributions having one to three parameters: Exponential, Weibull, Gamma, Truncated Lognormal, Loglogistic and Generalized Pareto, and generalized parametric distributions having three or more parameters: the generalized beta distribution of the second kind and the G-and-H distribution. The estimation of the parameters is based on some well-known techniques as maximum likelihood, method of moments, or quantile estimation [4, 21]. Other techniques in modelling severity distributions are based on Extreme Value Theory (EVT) and non-parametric empirical sampling. EVT is a branch of statistics to study the extreme phenomena (large operational losses, for example).

Empirical sampling simulates losses from the empirical distribution in the following manner. For each simulated period:

1. Draw a frequency N from the Poisson distribution (the annually aggregated number of losses shows a Poisson behaviour);
2. Generate N loss severities (with replacement) using the original set of loss severity data: L_i (i =.1, 2, …, N).
3. Sum the N losses to get the total annual loss: $S = \Sigma\{L_i \mid i = 1, 2, \ldots, N\}$.
4. Repeat the steps 2 and 3 for many times (usually thousands).

The distribution of S (according to LDA) is called the *aggregate loss distribution*, and the risk exposure can be measured as a quantile of S.

DFA is an important development based on stochastic simulation (Monte Carlo methods [12], [17], [21]) used mostly in non-life insurance and reinsurance [5]. However, for life insurance, the approach called Asset Liability Management (ALM) if using stochastic simulation becomes similar to DFA. It is accepted now that DFA is a variant of ALM, showing a greater emphasis on both economic scenario generators and the interrelationships between assets and liabilities.

Applying this approach any enterprise or financial company is able to investigate the potential impact during decisional process. According to [18], the DFA borrows many concepts and methods from economics and statistics and integrates them in a powerful tool (actually implemented in software as decision support systems). DFA requires a scenario generator and a calibration procedure. Also is mandatory a multivariate company model, an analysis and presentation module together with a control and optimization module for improving the strategy [5].

The *scenario generator* is a module which implements stochastic models for risk factors, affecting the company strategic decisions, like economic risks, liability risks, asset risks and business risks. On the first step in developing a DFA model it will be investigated the risks and the factors affecting the company results. Then, the objective functions and the projection period (the planning horizon, usually long enough) have to be chosen.

The *calibration procedure* is responsible for finding the suitable parameters of the models used in the scenarios generation. The difference between DFA and classical scenarios testing approach results from the usage of Monte Carlo simulation (including bootstrap) during scenarios generation and calibration.

According to [5], "the real challenge of DFA scenario generation lies in the composition of the component models into an integrated model". This task will be accomplished using both the deterministic modeling (functional approximation) and the statistical modeling (correlation, multivariate statistics, time series etc) [4, 7, 8, 9, 14]. When there is only a small data set, the estimation of the model parameters (in general belonging to a high-dimensional space) asks for uncertainty diminishing. The bootstrap approach can be used for accuracy estimation as described in [2, 6, 10, 11, 20]. After calibration, if necessary, the stochastic model can be reconsidered and improved.

The output of the scenario generator is a large number of Monte Carlo scenarios representing the future possible states of the company, while only a small set of scenarios are used in classical scenario testing approach. The company model has to react to the generated scenario according to the company operating structure along aspects like insurance, investment, the impact of reinsurance, under specific regulation, accounting and taxation. Company models used in DFA vary from simple to highly complex size. Usually, such models imitate the cash flows of the company (mainly the technical and financial accounting structures) in a risk management manner.

Running DFA analysis produces results for the output variables under interest along some future moments of time (predicted values). The size of the obtained set of results is, in general, huge enough to ask for data mining techniques used by the analysis procedure. The common approach in data mining for such task is based on statistical methods for data analysis. During analysis, the predicted values will be obtained with some accuracy to be assessed also by computer intensive methods like cross-validation or bootstrap [20]. It is possible that some scenarios will provide unacceptable results. This situation generates corrective measures and re-running the DFA. The control and optimization module of the DFA system is responsible for such a task. After the optimization step, the presentation module will provide reports obtained from the DFA. The management board will choose from these reports the best strategy for the company. Even "DFA models provide generally deeper insight into risk and potential rewards of business strategies than scenario testing can do", some weaknesses of the DFA there exist [18]. The first difficulty consists of capturing the complexity of the real-life business environment. The second one deals with the strong dependence of the results on the assumptions used in the model set-up. For small models the results can correspond with intuition. However, more effort is necessary to understand and control the uncertainties and approximations for a useful DFA approach.

3. **Operational loss data in environmental economics**

Windstorms are the most important extreme events causing severe damages. Breakage is the most common type of storm damage. The degree of impact depends on the degree and pattern of damage as well as the tree species involved. The cyclonic winds cause twisting and separation of wood fibbers in the main stem. The uprooted trees will be degraded quickly by stains, decays, and secondary insects such as Ips (engraver) bark beetles, borers, powder post beetles, and ambrosia beetles. Also, during storms, many trees sustain wounds caused by falling tops, adjacent uprooted trees, and major branch breakage. Other damages come from bent trees or standing water.

The impact has significance not only for environment saving, but also from business point of view, including insurance [4, 7, 8].

The Weibull density function was considered for describing the operational loss data given by [16] related to wind catastrophes that occurred in 1977:

$$\mathrm{f}\left(\mathrm{x};\alpha,\beta\right)=\frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}e^{-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} \tag{3}$$

where $x \geq \gamma$, $\alpha > 0$ (the shape parameter), $\beta > 0$ (the scale parameter), and $\gamma > 0$ (the position parameter). Let us denote by F(.) the corresponding cdf. The bootstrap method can be used to obtain the standard error of the estimated parameters, denoted by $\hat{\alpha},\hat{\beta}$ and $\hat{\gamma}$. If θ stands for any of the parameters $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$, then the approximation to the bootstrap estimate of standard error of θ is given by:

$$se\left(\hat{\theta}\right)=\frac{1}{\sqrt{B-1}}\left(\sum_{k=1}^{B}\left(\hat{\theta}^{*}\left(k\right)-\eta\right)^{2}\right)^{\frac{1}{2}} \tag{4}$$

where B is the total number of resampling simulations, and $\hat{\theta}^{*}(k)$ is the estimated version of the corresponding parameter during the resampling step indexed by k, and

$$\eta=\frac{1}{B}\sum_{k=1}^{B}\hat{\theta}^{*}\left(k\right). \tag{5}$$

For the data set considered, using $\gamma = 0$, we found, after one thousand resampling simulations, the next five shapes of the cdf, respective pdf, Weibull wind loss function presented in figure 1.

Also, after one thousand resampling simulations, the standard error for the parameters of the Weibull pdf was found as: se($\hat{\alpha}$) = 0.00164 and se($\hat{\beta}$) = 0.00115.

The threat of wind damage is widely recognized to be a serious constraint to the management of forests. An increase in the frequency of intense storms could potentially have a significant effect. Therefore, is possible to experience serial correlation in data, so in order to predict the changes in environment generated by forest damages, the bootstrap approach has to be used carefully. After some simulations of these cases, we recommend a segment type (block) resampling, not a standard resampling. The segment size has to be identified during a data mining process.

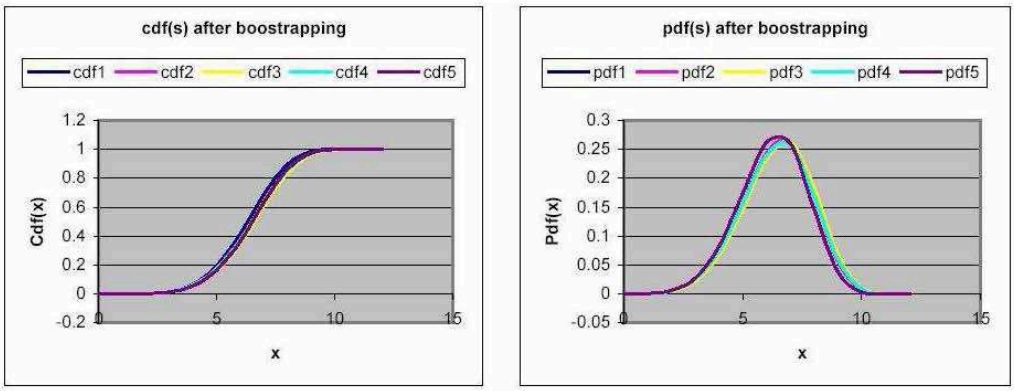

Figure 1. *Bootstrapping Weibull wind loss models*

4. **Bootstraping the expected return on portfolios**

In the following, an investigation on expected return on portfolios based on bootstrapping generalized regression will be presented. By portfolio, according to [21] we mean a group of financial assets. It is expected that "a rational investor" choose his/her portfolio so as to maximize the expected return and to minimize the risk.

Let us consider N assets, 1, 2,…, N, and the wealth equal to 1. The portfolio is modelled as the vector $x = (x_1, x_2, \ldots, x_N)^T$, where x_k represents the fraction of the unit wealth invested in the k^{th} asset, k = 1, 2,…, N, so that $x_1+x_2+\ldots+ x_N = 1$ [9, 15]. The returns of the N mentioned assets are random variables $\rho = (\rho_1, \rho_2, \ldots, \rho_N)^T$, with the expected returns $r = E\rho = (r_1, r_2, \ldots, r_N)^T$, and the covariance matrix $V = (v_{ij})$, where $v_{ij} = \text{cov}(\rho_i, \rho_j)$, i, j = 1, 2, …, N. As, usually, let us denote the diagonal elements by $\sigma_i^2 = v_{ii}$, with σ_i being the standard deviations of the returns.

For a given portfolio P, represented by weights x, the expected return on the portfolio P is $r_P = r^T x$ and the variance of the portfolio P is $\sigma_P^2 = x^T V x$, while the risk of the portfolio P is σ_P.

Usually, the return on an asset is explained in terms of a linear combination of more factors:

$$\rho_i = \sum_{j=1}^{m} \beta_{ij} F_j + \varepsilon_i,\ i = 1, 2, \ldots, N \tag{6}$$

where F_j (j = 1, 2, …,m) are observed explanatory variables (like production, inflation, term structure and other economic factors), ε_i is a zero mean random disturbance (not observable), and (β_{ij}) are unknown parameters which are specific for the given asset. If there are T observations (usually gathered historical data), the regression model is

$$\rho_i = F^T \beta_i + \varepsilon_i, \tag{7}$$

where $\rho_i = (\rho_{i1}, \rho_{i2}, ..., \rho_{iT})^T$, $F := (F_{ij})$, $i = 1,2,...,m$ and $j = 1,2,...,T$, is an m x T matrix, $\beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{im})^T$, and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{iT})^T$, with $E\varepsilon_{ij} = 0$ and $\mathrm{cov}(\varepsilon_{ik}, \varepsilon_{il}) = \sigma_i^2 \delta_{kl}$, $k,l = 1,2,...,T$, where $\delta_{kl} = 1$ for $k = l$ and $\delta_{kl} = 0$, otherwise.

For T > m and rank (F) = m, the ordinary least square estimator of β_i is

$$\hat{\beta}_i = (FF^T)^{-1} F\rho_i \tag{8}$$

having the covariance matrix

$$\mathrm{cov}(\hat{\beta}_i) = \sigma_i^2 (FF^T)^{-1} \tag{9}$$

while an unbiased estimator for σ_i^2 is given by

$$\sigma_i^2 = \frac{1}{T-m-1}(\rho_i - F^T\hat{\beta}_i)^T (\rho_i - F^T\hat{\beta}_i). \tag{10}$$

However, when T < m (small data set; short history), or rank (F) < m, then a least square estimator of β_i is computed using the generalized inverse [3] of F^T:

$$\hat{\beta}_i = (F^T)^+ \rho_i. \tag{11}$$

In this case, the confidence intervals for β_i will be obtained by bootstrapping, the generalized least square estimate being

$$\hat{\beta}^*_i = (F^T)^+ \rho^*_i. \tag{12}$$

where

$$\rho^*_i = F^T\hat{\beta}_i + \varepsilon^*_i, \tag{13}$$

and ε^*_i is obtained by resampling the centred residuals using the empirical distribution.

For a moderate number of bootstrap steps (about 20), the statistics concerning the regression model will be obtained by analysing the bootstrap results.

Let us consider some data related to the return on the investment (RET) in forest industry when damages generated by windstorms appear. There are both vulnerable and resistant regions [7]. Let F_1 be the initial investment, F_2 be the annual growth rate, F_3 be the affected surface of vulnerable region, F_4 be the vulnerable surface, F_5 be the rate of declassification of the vulnerable region, F_6 be the affected surface of resistant region, F_7 be the resistant surface and F_8 be the rate of declassification of the resistant region. For twelve classes of regions, a regression analysis was done, and after twenty bootstrap resampling steps the following results where obtained:

$$RET = 1.013F_1 + 0.115F_2 + 5.1E-0.2F_3 + 6.1E-0.3F_4 - \\ -15.58F_5 + 0.15F_6 - 8.5E-0.3F_7 + 13.71F_8 ,$$

with the following 95% confidence intervals for the coefficients: (0.906, 1.119), (-0.627,0.859), (-0.411,0.513), (-0.083, 0.095), (-14.102, 115.951), (-0.262, 0.556), (-0.059, 0.042) and (-116.435,143.862). The present model explains more than 82% from the data variability, the proportion of variability being 0.998.

5. Conclusions

This paper describes the usage of some bootstrapping techniques to be used in business statistics, mainly concerning dynamic financial analysis and loss distribution functions estimation in the framework of the environmental economics. The investigation outlines an important number of opportunities for bootstrapping in financial and environmental data analysis.

References

1. Albeanu, G., Popentiu, F., *On the Bootstrap Method: Software Reliability Assessment and Simultaneous Confidence Bands*, Annals of Oradea University, Energetic Series, 7(1), 109-113, 2001.
2. Albeanu, G., *Resampling Simultaneous Confidence Bands for Nonlinear Explicit Regression Models*, Mathematical Reports, 50(5-6), 289-295, 1998.
3. Ben-Israel, A., Greville, T.N.E., *Generalized Inverses: Theory and Applications*, Wiley, New York, 1977.
4. Berthouex, P.M., Brown, L.C., *Statistics for Environmental Engineers*, CRC Press LLC, Boca Raton, 2002.
5. Blum, P., Dacorogna, M., *DFA-Dynamic Financial Analysis*, in Encyclopaedia of Actuarial Scienc, John Wiley & Sons, 2004.
6. Derrig, R.A., Ostaszewski, K.M., Rempala, G.A., *Applications of Resampling Methods in Actuarial Practice*, Proceedings of Casualty Actuarial Society, 87, 222-264, 2000.
7. Drăgoi, M., *Forest Economics* (Economie forestieră), Editura Economică, Bucharest, 2000 (in Romanian).
8. Drăgoi, Simona, Albeanu, G., *Markov Chain Prognosis Model for Allowable Cut Structure*, Lesnitvi- Forestry, 44(8), 344-347, 1998.
9. Dupacová, J., Hurt, J., Štepán, J., *Stochastic Modeling in Economics and Finance*, Applied Optimization Vol. 75, Kluwer, New York, 2002.
10. Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
11. Efron, B., *Computer-Intensive Methods in Statistical Regression*, SIAM Review, 30(3), 421-449, 1988.
12. Fishman, G., Monte Carlo: *Concepts, Algorithms, and Applications*, Springer, Berlin, 1996.
13. Freedman, D.A., *Bootstrapping Regression Models*, The Annals of Statistics, 9(6), 1218-1228, 1981.
14. Ghica, M., *A Risk Exchange Model with a Mixture Exponential Utility Function*, Annals of Bucharest University, Mathematics and Informatics Series, 2006.
15. Ghica, M., *Optimal Portfolios in Finance and Actuarial Science*, Annals of *Spiru Haret* University, Mathematics and Informatics Series, 2006 (in Romanian).
16. Hogg, R.V., Klugman, S.A., *Loss Distributions*, John Wiley & Sons, New York, 1984.
17. Jackel, P., *Monte Carlo Methods in Finance*, John Wiley & Sons, Chichester, 2002.

18. Kaufmann, R., Gadmer, A., Kett, R., *Introduction to Dynamic Functional Analysis*, ASTIN Bulletin, 31(1), 213-249, 2001.
19. Klugman, S.A., H.H. Panjer, H.H., Willmot, G.E., *Loss Models: From Data to Decisions*, John Wiley & Sons, Inc., New York, 1998.
20. Kohavi, R., *A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection*, International Joint Conference on Artificial Intelligence (IJCAI), August 20-25 1995, Morgan Kaufmann,Vol. 2, pp. 1137-1145, Montréal, Québec, Canada, 1995.
21. Malliaris, A.G., Brock, W.A., *Stochastic Methods in Economics and Finance*, Elsevier Science B.V., Amsterdam, 1982.
22. Ostazsewski, K.M., Rempala, G.A., *Parametric and Nonparametric Bootstrap in Actuarial Practice*, The Actuarial Foundation, 2000.
23. Pallini, A., *Efficient Bootstrap Estimation of Distribution Functions*, METRON, 58(1-2), 81-95, 2000.
24. Rachev, S.T., Chernbai, A. and Menn, C., *Empirical Examination of Operational Loss Distributions*, Karlsruhe University Technical Report, 2005.
25. Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.
26. Web phrases, *Meanings and Origins of Phrases, Sayings and Idioms*, http://www.phrases.org.uk/meanings/290800.html, 2007.