STATISTICAL MODELLING
P.G.M. van der Heijden, W. Jansen, B. Francis and G.U.H. Seeber (editors)
1992 Elsevier Science Publishers B.V.

Bootstrap Importance Sampling in Regression

Anna Gigli

Istituto per le Applicazioni del Calcolo, viale del Policlinico 137, 00142 Rome, Italy

Keywords: Generalized linear model; non-nested models; deviance.

1. INTRODUCTION

Efficient bootstrap methods have been intensively studied over the past few years, although almost exclusively with statistics based on simple random samples. The purpose of this paper is to investigate the use of importance sampling as an efficient bootstrap method for the analysis of regression models. We consider parametric bootstrap which involves simulation from a fitted parametric model. In Section 2 we give efficiency calculations for the parametric bootstrap applied to the normal linear model. Section 3 applies the parametric bootstrap to the generalized linear model, with efficiency calculations and Section 4 describes some illustrative examples involving non-nested models.

2. LINEAR REGRESSION

The concept of *importance sampling* was first introduced by Hammersley & Morton (1956), in an article on Monte Carlo methods. They describe it as a way of concentrating the distribution of the sample points in that region of the sample space of the most "importance"; in other words there is a deliberate alteration of the sampling probabilities in order to reduce the variability of some estimate of interest.

This idea can be used to make the bootstrap more efficient and cost-effective: if, instead of resampling uniformly from $\hat{F}$ (an appropriate estimate of the distribution function F), we resample from a distribution which is more concentrated around the parameter we want to estimate, the variation of the new estimate should be smaller than the variation of the classical bootstrap estimate, and this would allow us to decrease the number of simulations.

Our basic framework is the following:

Let $y_1, \ldots, y_n$ be n realizations of the random variable Y, and $x_1, \ldots, x_n$ be the corresponding values of the covariate x. Suppose we fit two regression models to the data:

$H_0 : \; y_i = \eta_0 + \epsilon_i ,$

$H_1 : \; y_i = \eta_1 + \epsilon_i ,$

where $\eta_{0i} = \beta_0 , \eta_{1i} = \beta_0 + \beta x_i , \epsilon_i$'s are iid with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2 < \infty$, but nothing else is known about their common distribution F. We are interested in comparing the two models, i.e. in assessing whether β should be zero or not. Under H_1, using the method of least squares, we estimate β as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

with $E_1(\hat{\beta}) = \beta$ and $var_1(\hat{\beta}) = \sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2$. [From now on we will refer to the expectation, variance, etc. under the hypothesis H_i as to E_i, var_i, etc.]. We choose the test statistic $T = \hat{\beta}/s.e.(\hat{\beta})$, and aim to estimate the tail probability of T under H_0: $p_{obs} = pr(T \geq t_{obs} \mid H_0)$, where t_{obs} is a known value of T. Because of the unknown nature of the distribution F we cannot estimate p_{obs} directly and will make use of a Monte Carlo method. In order to illustrate and verify our proposed method, we will apply it to the case where the distribution F is known to be normal with mean zero and variance σ^2, and then we compare the approximate results obtained with the exact results provided by the normal theory. Under normal assumptions, if $\beta = 0$ (i.e. H_0 holds) and σ^2 is known, we know that $p_{obs} = pr(T \geq t_{obs} \mid H_0) = 1 - \Phi(t_{obs})$.

Efron (1982) proposes a bootstrap method for regression which consists of bootstrapping the residuals of the regression model:

let $T^* = \hat{\beta}^*/s.e.(\hat{\beta}^*) = \sum_{i=1}^{n}(x_i - \bar{x})y_i^*/\sigma\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ be the statistic of interest calculated from the simulated distribution, where y^* is simulated from the null model $y_i^* = \bar{y} + \epsilon_i^*$ and the ϵ_i^*'s are uniformly sampled from the set of the residuals $\hat{e}_i = y_i - \hat{y}_i$. A bootstrap estimate of the tail probability p_{obs} is $\hat{p}_0 = (1/B)\sum_{b=1}^{B} I(T_b^* \geq t_{obs})$, where $t_{obs} = \hat{\beta}/s.e.(\hat{\beta})$ is obtained from the data, $I(x) = 1$ if $x \geq 0$, or $= 0$ otherwise is the indicator function, and B is the number of independent bootstrap replications.

Since $E_0^*(\hat{p}_0) = (1/B)\sum_{b=1}^{B} I(T_b^* \geq t_{obs}) = pr^*(T^* \geq t_{obs})$, it follows that asymptotically $E_0^*(\hat{p}_0) = p_{obs}$ and $var_0^*(\hat{p}_0) = (p_{obs} - p_{obs}^2)/B$, where E_0^* and var_0^* are the bootstrap expectation and variance calculated with respect to the null model.

Although $\hat{p}_0$ is an unbiased estimate of p_{obs}, when the event $\{T^* \geq t_{obs}\}$ does not happen too often (for instance if t_{obs} is in the tail of the distribution), the variance of $\hat{p}_0$ can be big, unless a large number of bootstrap simulations (typically of the order of 1,000) is used.

Alternatively, as suggested by Johns (1988), we can use the importance sampling bootstrap which consists of generating a set of *iid* Y^*'s *no longer from the null distribution* $F_0(\hat{\eta}_0)$, but from a suitable $F_\alpha(\hat{\eta}_\alpha)$ and then downweighting the events $\{T^* \geq t_{obs}\}$ by the ratio of the two likelihoods evaluated at $\boldsymbol{Y}^*$. We obtain:

$$\hat{p}_\alpha = \frac{1}{B}\sum_{b=1}^{B}\left[I\{T_b^* \geq t_{obs} \mid Y^* \sim F_\alpha(\hat{\eta}_\alpha)\}\prod_{i=1}^{n}\frac{f_0(Y_i^*)}{f_\alpha(Y_i^*)}\right],$$

where we assume that $f_0 = F_0'$ and $f_\alpha = F_\alpha'$ represent the corresponding densities.

The mean and variance of $\hat{p}_\alpha$ under the model F_α are determined by

$$\begin{aligned} E_\alpha^*(\hat{p}_\alpha) &= \int I(T^* \geq t_{obs})\frac{f_0(z)}{f_\alpha(z)}f_\alpha(z)dz \\ &= \int I(T^* \geq t_{obs})f_0(z)dz = E_0(\hat{p}_0) = p_{obs} \end{aligned} \tag{1}$$

$$
\begin{aligned}
E^*_\alpha(\hat{p}^2_\alpha) &= \int I(T^* \geq t_{obs}) \frac{f_0^2(z)}{f_\alpha^2(z)} f_\alpha(z) dz \\
&= \int I(T^* \geq t_{obs}) \frac{f_0(z)}{f_\alpha(z)} f_0(z) dz = p_{is} \,,
\end{aligned} \tag{2}
$$

say, and therefore $var^*_\alpha(\hat{p}_\alpha) = (p_{is} - p^2_{obs})/B$.

If we choose the model F_α in some appropriate way we can find a more efficient estimator and this will allow us to reduce the number of bootstrap simulations. Let

$$
eff(\hat{p}_0\,;\,\hat{p}_\alpha) = \frac{var^*_0(\hat{p}_0)}{var^*_\alpha(\hat{p}_\alpha)} = \frac{p_{obs} - p^2_{obs}}{p_{is} - p^2_{obs}} \tag{3}
$$

denote the theoretical gain in efficiency of $\hat{p}_\alpha$ over $\hat{p}_0$; $eff(\hat{p}_0\,;\,\hat{p}_\alpha)$ will be maximum when p_{is} is at a minimum. Hence we are interested in finding the value of α that minimizes p_{is}. For a reason that will become apparent later, α is called the *tilting coefficient*.

In order to find the optimum α we provide an explicit form for η_0 and η_α: we assume that η_0 is a straight horizontal line and η_α a line of slope α:

$$
\eta_0 = \bar{y} \;,\; \eta_\alpha = \bar{y} + \alpha(x - \bar{x}) \,,
$$

and that the variance of the two distributions F_0 and F_α is the same σ^2 (in fact our initial calculations considered two different variances, but later we found that the simplified version produces results nearly as optimal).

This produces a simplified version of p_{is}:

$$
p_{is}(\alpha, \sigma) = \exp\left(\frac{\alpha^2}{var(\hat{\beta})}\right)\left\{1 - \Phi\left(\frac{\alpha + \hat{\beta}}{\sqrt{var(\hat{\beta})}}\right)\right\} ,
$$

where $var(\hat{\beta}) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$.

If we make use of the expansion $1 - \Phi(u) = \phi(u)/u$ (see, for example, Barndorff-Nielsen & Cox, 1989, pp.55-57), the optimum value for α obtained from the corresponding p_{appr} will simply be

$$
\tilde{\alpha} = \sqrt{\hat{\beta}^2 + var(\hat{\beta})} \,.
$$

For this optimum value of α, we obtain

$$
p_{is}(\beta) = \exp\left(\frac{\hat{\beta}^2 + var(\hat{\beta})}{var(\hat{\beta})}\right)\left\{1 - \Phi\left(\frac{\sqrt{\hat{\beta}^2 + var(\hat{\beta})} + \hat{\beta}}{\sqrt{var(\hat{\beta})}}\right)\right\} .
$$

Setting $t = \hat{\beta}/\sqrt{var(\hat{\beta})}$, we have:

$$
p_{is}(t) = \exp(t^2 + 1)\{1 - \Phi(\sqrt{t^2 + 1} + t)\} .
$$

The table below illustrates the gain in efficiency for a few values of t. It confirms what we said previously, that is importance sampling works well for assessing probabilities for rare events: as we move towards the tail of the distribution the gain in efficiency grows

Table 1: Gain in efficiency as function of t

t	p_{is}	p_{appr}	eff
-1	2.51	6.53	0.07
-0.5	0.94	1.86	0.47
-0.4	0.79	1.49	0.62
-0.1	0.50	0.80	1.18
0	0.43	0.66	1.38
0.1	0.37	0.54	1.57
0.5	0.18	0.23	2.39
1	0.06	0.07	4.03
2	0.002	0.002	19
3	0.8×10^{-5}	0.8×10^{-5}	222
4	5.4×10^{-9}	5.6×10^{-9}	7263

exponentially, but if we remain towards the centre of the distribution we will obtain no substantial gain. That explains also the lack of symmetry in eff as a function of t, even though t is symmetric.

3. GENERALIZED LINEAR MODELS

Let Y be a r.v. with density $f(y;\eta,\phi)$:

$$f(y;\eta,\phi)=\exp\left\{\frac{y\eta-b(\eta)}{a(\phi)}+c(y;\phi)\right\} \tag{4}$$

and $\eta=g(\mu)$ be the link function, according to the terminology illustrated by McCullagh & Nelder (1989).

Given a vector $\boldsymbol{y}$ of realizations of $\boldsymbol{Y}$, we want to choose between two models to fit to the data:

$$H_0 : y\sim F_0(\eta_0)$$
$$H_1 : y\sim F_1(\eta_1),$$

where F_0 and F_1 are two distributions belonging to the family (4).

Let $T=D_0(y;\mu_0)-D_1(y;\mu_1)$ be the difference of deviances between the two models, as in McCullagh & Nelder (1989). For convenience we take H_0 to be the hypothesis producing the largest deviance, so that $T>0$.

Our aim is to find a significance level for H_0, ie to estimate

$$p_{obs}=Pr(T\geq t_{obs}\mid H_0),$$

where t_{obs} is the observed value of T when the two models are fitted.

In absence of standard methods of estimating p_{obs} (if the two models are non-nested), one way would be via a bootstrap procedure:

generate a set of *iid* y^*'s from $F_0(\hat{\eta}_0)$, calculate $T^* = D_0(y^*; \hat{\mu}_0^*) - D_1(y^*; \hat{\mu}_1^*)$, repeat the same procedure B times and calculate the bootstrap significance level $\hat{p}_0 = (1/B)\sum_{b=1}^{B} I(T_b^* \geq t_{obs} \mid y^* \sim F_0(\hat{\eta}_0))$.

As in the regression case, $E_0^*(\hat{p}_0) = p_{obs}$ and $var_0^*(\hat{p}_0) = (p_{obs} - p_{obs}^2)/B$.

In a similar fashion to the normal case, we find that the optimum $\tilde{\alpha}$ for p_{is} is the solution

$$\tilde{\alpha} = \sqrt{\left[\frac{\sum_{i=1}^{n}\{y_i - b'(\hat{\eta}_{0i})\}(\hat{\eta}_{0i} - \hat{\eta}_{1i})/a_i(\phi)}{\sum_{i=1}^{n} b''(\hat{\eta}_{0i})(\hat{\eta}_{0i} - \hat{\eta}_{1i})^2/a_i(\phi)}\right]^2 + \frac{1}{\sum_{i=1}^{n} b''(\hat{\eta}_{0i})(\hat{\eta}_{0i} - \hat{\eta}_{1i})^2/a_i(\phi)}}\,.$$

[We refer to a chapter in the PhD thesis by Gigli for details on the calculations].

4. EXAMPLES

In the following examples we concentrate our attention on two kinds of non-nested models: in the first two cases we assume a specific error distribution and want to compare two different link functions. In the final example we compare two different error distributions. The following bootstrap method provides an alternative to the parametric approach suggested by many (see for example Cox, 1961; Atkinson, 1979; Davison & Gigli, 1989; and Pereira, 1977 for a review of the non-nested models literature).

4.1 POISSON MODELS

Wahrendorf *et al.*(1987) propose a bootstrap method to analyze the goodness-of-fit of different models to death rate from coronary heart disease among British male doctors. The aim of the fitting was to investigate whether an additive or multiplicative model best explains the influence of age and smoking habits on the death rate (see Breslow, 1985 for a full description of the data). Two GLM's with Poisson error and respectively additive (A) and multiplicative (M) links are fitted to the data, resulting in a deviance of 13.05 with 3 df for the multiplicative model, and 3.37 (with 3 df) for the additive model.

In their approach Wahrendorf *et al.*consider the difference of deviances as the test statistic ($t_{obs} = 9.674$), to test the hypothesis that the two models fit the data equally well. Their bootstrap procedure consists in simulating bootstrap r.v.'s y^*'s from the original data (ie $\hat{\mu}_i$, the fitted mean of the Poisson distribution is taken to be y_i, the corresponding observed datum), and fitting the two models to the y^*'s; the bootstrap test statistic is $T^* = D(y^*; \hat{\mu}_M) - D(y^*; \hat{\mu}_A)$, where M and A stand for multiplicative and additive and they construct the bootstrap distribution of T^*; finally they calculate a bootstrap confidence interval and if it contains the value zero, they accept the hypothesis that the two models explain the data equally well.

But a few problems arise: i) simulating bootstrap samples from the data does not seem correct, as the simulation should be done from a null hypothesis related to the models; in other words, in the null hypothesis one should state the preference towards one model or the other, and from there infer whether the two models describe the data equally well, or if one of the two is to be preferred; the result of their test is conservative not because the variance of the test is too large (this is their explanation), but because the sampling is done from the wrong set; ii) the number of simulations required is quite large - 1,000 for this example.

In the importance sampling framework the approach is quite straightforward: since the additive model provides a smaller deviance, we simulate the bootstrap samples from a Poisson distribution with mean vector $\hat{\boldsymbol{\mu}}_A$, which represents for us the alternative hypothesis, and then downweight the corresponding observation by the likelihood ratio of the null over the alternative hypotheses, obtaining an estimate of the significance level of the null hypothesis (the multiplicative model)

$$\hat{p}_{M,is} = \frac{1}{B} \sum_{b=1}^{B} \left[I\{T_b^* \geq t_{obs} \mid \boldsymbol{y}^* \sim P(\hat{\boldsymbol{\mu}}_A)\} \prod_{i=1}^{n} \frac{f_M(y_i^*)}{f_A(y_i^*)} \right]. \tag{5}$$

The result of $B = 200$ bootstrap simulations is $\hat{p}_{M,is} = 0.00064$ (se $= 0.000096$).

As a comparison, we run a standard bootstrap experiment (in which the bootstrap sample is drawn from the null hypothesis, the multiplicative model)

$$\hat{p}_{M,cf} = \frac{1}{B} \sum_{b=1}^{B} I\{T^* \geq t_{obs} \mid \boldsymbol{y}^* \sim P(\hat{\mu}_M)\}.$$

As suggested by the value of $\hat{p}_{M,is}$, the event $T^* \geq t_{obs}$ has a very low frequency under H_0: it will happen approximately 6.4 times out of 10,000, and therefore a large number of bootstrap replicates is needed in order to have a non-zero estimate; $B = 5,000$ bootstrap simulations were performed, and the result is $\hat{p}_{M,cf} = 0.00080$ (se $= 0.000400$).

The two estimates are consistent but the standard error of the standard estimate is more than 4 times higher than that of the importance sampling estimate. The theoretical gain in efficiency of the importance sampling estimate over the standard estimate is given by (3) and we approximate it by $\widehat{eff} = (\hat{p}_{M,is} - \hat{p}_{M,is}^2)/\widehat{var}(\hat{p}_{M,is}) = 349$.

Notice that at no extra cost we obtain also an estimate of the significance level for the alternative hypothesis: from (5), if we ignore the weights, we have

$$\hat{p}_{A,cf} = \frac{1}{B} \sum_{b=1}^{B} I\{T_b^* \geq t_{obs} \mid \boldsymbol{y}^* \sim P(\hat{\boldsymbol{\mu}}_A)\} = 0.74 \text{ (se} = 0.031).$$

The conclusion we draw from the test is different from that of Wahrendorf *et al.*: as the estimated significance level corresponding to the multiplicative model is very low, we infer that the additive model fits the data better.

We could attain an even smaller variance of our estimate of p_{obs}, if we sampled from a Poisson($\hat{\mu}_\alpha$) distribution, with α suitably chosen. The optimum tilting parameter $\tilde{\alpha}$ in this case is equal to 0.75. If $\hat{\eta} = \eta(\hat{\mu})$ is the fitted canonical linear predictor, then $\tilde{\alpha}$ provides us with the optimum $\hat{\boldsymbol{\eta}}_{\tilde{\alpha}} = 0.75\, \hat{\boldsymbol{\eta}}_A + 0.25\, \hat{\boldsymbol{\eta}}_M$. The estimate of the significance level of H_0 obtained by tilting the importance sampling distribution by $\tilde{\alpha}$ is $\hat{p}_{M,\alpha} = 0.0005493$ (se $= 0.000080$) with a gain in efficiency of $\hat{p}_{M,\alpha}$ over $\hat{p}_{M,is}$ of $\widehat{eff}(\hat{p}_{M,\alpha}; \hat{p}_{M,is}) = 1.39$.

Fig. 1 is a plot of the deviance for different values of the linear predictor, where $\hat{\boldsymbol{\eta}}_A$ corresponds to $\alpha = 1$ and $\hat{\boldsymbol{\eta}}_M$ to $\alpha = 0$. The minimum deviance lies somewhere in between 0.75 and 1, which is a further confirmation for the additive model.

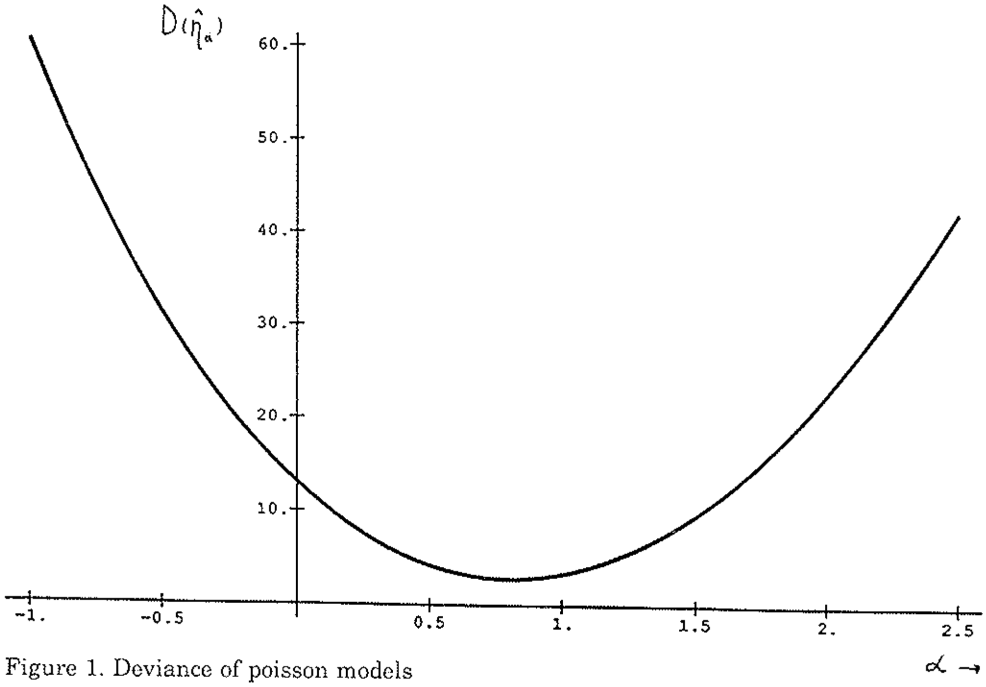

Figure 1. Deviance of poisson models

4.2 BINOMIAL MODELS

We consider the data analyzed by Prentice (1976) on the mortality of adult beetles after exposure to a poisonous gas: the counts of beetles killed are classified according to different dosages of the gas and are modelled as binomial r.v.'s, with the number of beetles exposed as denominators. The two models we want to compare are the *logit* (L), and the *complementary log-log* (C).

The deviances of the two fitted models are $D(y;\, \hat{\mu}_L) = 11.232$, and $D(y;\, \hat{\mu}_C) = 3.4464$ (i.e. $t_{obs} = 7.786$), and we want to find an estimate of p_{obs}, the significance level of the null hypothesis H_0: logistic model.

The employment of the importance sampling and standard bootstrap procedures to estimate p_{obs} leads to the following results: $\hat{p}_{L,is} = 0.0031$, with s.e. $= 0.000385$, on B $=$ 200 simulations, and $\hat{p}_{L,ef} = 0.0040$, with s.e. $= 0.001997$, on B $= 1{,}000$ simulations, and estimated theoretical gain in efficiency $\widehat{eff} \approx 104$.

This result, together with the estimated significance level of the alternative model ($\hat{p}_{C,ef} = 0.46$, with s.e. $= 0.0353$, on 200 simulations) tells us that between the logistic and the complementary log-log models the latter fits the data much better. (The fact that the logistic model was unsuitable had already been stated by Pregibon, 1980).

It is interesting to notice the discrepancy between the variance of $\hat{p}$ when the sampling is done from the alternative (complementary log-log) model and the variance of $\hat{p}_{\hat{\alpha}}$ (i.e. when the sample is drawn from $f_{\hat{\alpha}}$), where $\hat{\alpha}$ is the α that theoretically gives the largest reduction in variance. According to our theory, the minimum variance should be attained for $\hat{\alpha} = 0.55$. However, the simulation study tells us that the minimum variance is attained at the alternative model, that is when $\alpha = 1$. Moreover the minimum deviance is also achieved at the alternative model.

The reason for that is that we based our theoretical calculations on the assumption that our test statistic is approximately normal and evidently in this specific problem the assumption does not hold. We therefore suggest simulating from the alternative distribution F_1 anyway, because it is "safer" than simulating from the (theoretically) optimum distribution $F_{\hat{\alpha}}$.

4.3 LOGNORMAL VERSUS GAMMA MODELS

When we are interested in comparing two probability models that do not belong to the same family, the use of the deviance test is inappropriate.

Let $F_0(\boldsymbol{y};\boldsymbol{\mu}_0,\phi_0)$ and $F_1(\boldsymbol{y};\boldsymbol{\mu}_1,\phi_1)$ be the two models of interest. If we fit the two models to the data $\boldsymbol{y}$, the difference of deviances will be $D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0,\phi_0) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1,\phi_1)$ which is different from $2\{l(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1,\phi_1) - l(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0,\phi_0)\}$, the statistic we were using before, because the log-likelihoods of the two saturated models do not cancel out.

In this case a maximum likelihood ratio test appears to be more appropriate:

$$T = l(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1,\hat{\phi}_1) - l(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0,\hat{\phi}_0)\,,$$

where $\hat{\boldsymbol{\mu}}_0$, $\hat{\phi}_0$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\phi}_1$ are the m.l.e.'s of the respective parameters.

We applied this test to a data set on the cost of construction of nuclear power plants in USA, reported by Cox & Snell(1981, p.82). The data consists of the cost of 32 light water reactor power plants, and the objective of the analysis is to predict the capital cost involved in the construction of further similar power plants. We fit the models L = lognormal and G = gamma with 6 quantitative covariates (see Cox & Snell for a more extensive analysis of the problem), and we are interested in estimating the significance level of the null hypothesis $p_{obs} = pr(T \geq t_{obs} \mid L)$, where $t_{obs} = 0.2930$.

According to the importance sampling bootstrap, we have $\hat{p}_{L,is} = 0.1430$ with $s.e. = 0.02192$, on $B = 200$ simulations, while the Efron bootstrap gives us $\hat{p}_{L,ef} = 0.1400$ with $s.e. = 0.02460$, on $B = 200$ simulations, giving a gain in efficiency of the importance sampling method over the classical Efron method of $\widehat{eff} \approx 1.26$.

On the other hand, looking at the significance level of the alternative model: $\hat{p}_{G,ef} = 0.2100$ with $s.e. = 0.02887$, on $B = 200$ simulations, we can conclude that both models are plausible, though the gamma model appears to fit slightly better.

The insignificance of the gain in efficiency attained by the importance sampling bootstrap is due to the fact that we are not estimating an extreme tail probability (in fact the significance level of the lognormal hypothesis is about 0.14), and therefore we cannot expect an improvement when we implement the importance sampling bootstrap instead of the Efron bootstrap. This result stresses the fact that the importance sampling procedure is really effective in gaining efficiency only when examining a situation in which the event of interest occurs rarely.

Acknowledgements

This paper is part of a PhD project carried out at Imperial College, London, under the supervision of dr. A. Davison, to whom goes my gratitude for the valuable help. The research was supported by the Italian *Consiglio Nazionale delle Ricerche*.

References

Atkinson, A.C. (1970) A method for discriminating between models (with discussion). *J. R. Statistic. Soc.*, **32**, 323–353.

Barndorff-Nielsen, O. & Cox, D.R. (1989) *Asymptotic Techniques for Use in Statistics.* Chapman & Hall, London.

Breslow, N. (1985) Cohort analysis in epidemiology. In *A Celebration of Statistics*, 109–143, Springer, New York.

Cox, D.R. (1961) Test of separate families of hypotheses. *Proc. 4th Berkeley Symp.*, **1**, 105–123.

Cox, D.R. & Snell, E.J. (1981) *Applied Statistics - Principles and Examples.* Chapman & Hall, London.

Davison, A.C. & Gigli, A. (1989) Deviance residuals and normal scores plots. *Biometrika*, **76**, 211–212.

Efron, B. (1982) *The jackknife, the bootstrap, and other resampling plans.* CBMS Regional Conference Series in Applied Mathematics, **38**, SIAM Publications, Philadelphia.

Hammersley, J.M. & Morton, K.W. (1956) A new Monte Carlo technique: antithetic variates. *Proc. Camb. phil. Soc.*, **52**, 449–475.

Johns, M.V. (1988) Importance sampling bootstrap confidence intervals. *J. Amer. Stat Assoc.*, **83**, 709–714.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models.* Chapman & H London, 2 edition.

Pereira, B. de B. (1977) Discriminating among separate models: a bibliography. *Statist. Rev.*, **45**, 163–172.

Pregibon, D. (1980) Goodness of link test for generalized linear models. *Appl. Stati* **29**, 15–24.

Prentice, R.L. (1976) Generalization of probit and logit methods for dose-response cur *Biometrika*, **32**, 761–768.

Wahrendorf, J., Becher, H. & Brown, C.C. (1987) Bootstrap comparison for non-nes generalized linear models: applications in survival analysis and epidemiology. *A Statist.*, **36**, 72–81.