1.06 Resampling and Testing in Regression Models with Environmetrical Applications

J. Roca-Pardiñas, University of Vigo, Vigo, Spain

C. Cadarso-Suárez and W. González-Manteiga, University of Santiago de Compostela, Santiago de Compostela, Spain

1.06.1 Introduction to Bootstrap

The software revolution that has taken place in recent years has led to the development of different statistical methodologies of data analysis that rely on computer-based calculation. Preeminent among these techniques are the so-called resampling methods, which consist of generating a large number of samples to study the behavior pattern of given statistics. There are various procedures for generating artificial samples on the basis of an initial sample. Possibly the best known of these is the bootstrap method introduced by Efron in 1979,[1] a type of procedure that uses simulation to assess statistical accuracy. At present, bootstrap methods can be regarded as a general tool for statistical work, tending to be used in combination with other statistical techniques rather than in isolation. These types of methods are applied to different statistical areas, including the construction of confidence intervals (CIs), testing of hypotheses, and regression or principal components analysis, among others (see, for instance, the monograph by Efron and Tibshirani[2]). The interest that bootstrap methodology has aroused among the statistics community is reflected in the considerable number of textbooks devoted to justifying its theoretical bases or to discussing its applications in specific areas such as biology, environment, or medicine.[2–6] In addition, various packages have been developed by implementing a variety of bootstrap methods applied to different statistical methods, and most of these packages are in languages such as Fortran, R, or S-plus.

In many practical situations, the goal of research is to make inferences about a given characteristic (or parameter), θ, of a variable of interest, X. Estimation of this parameter is obtained on the basis of a statistic, $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, calculated as a function of a sample of size n $(X_1, \ldots, X_n)$, for example, a statistical inference about the mean population $\theta = \mu$ is based on the sample mean statistic $\hat{\theta} = \bar{X} = \sum_{i=1}^{n} X_i/n$.

A good part of conventional statistics is based on laws that enable the sample distribution of a statistic $\hat{\theta}$ to be approximated for sufficiently large sample sizes, e.g., where n is large enough, then, in accordance with the central limit theorem, the distribution of sample mean $\bar{X}$ approximately follows the distribution $N(\mu, S/\sqrt{n})$, where $S = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2/n}$ is the standard sample deviation. For instance, if one were seeking to construct a 95% CI for the population mean, this would be given by $\left(\bar{X} \mp 1.96S/\sqrt{n}\right)$.

Procedures of this type based on an asymptotic approach may display certain limitations in practice, as their performance will depend on the information furnished about the population by the sample. Hence, when not enough data are available, a good approximation of the distribution of $\hat{\theta}$ will not be obtained, and so the asymptotic method will not produce good results. At other times, moreover, one has to work with statistics that have no asymptotic laws governing their sample distribution. This is the case, for example, of difference in sample medians, or of sample asymmetry the sampling distribution of which is unknown. In the face of these types of limitations, various alternative procedures that enable the sample distribution of the target statistic to be obtained have emerged, by simulating a large number of random samples directly constructed on the basis of initially observed data. Among these resampling techniques is the bootstrap method, which is described in detail below.

As an alternative to the asymptotic method, the simplest version of the bootstrap method approximates the distribution of the statistic $\hat{\theta} = T(X_1, \ldots, X_n)$, in accordance with the following procedure:

1. Based on sample $\mathbf{X} = X_1, \ldots, X_n$, a random sample $\mathbf{X}^{*b} = X_1^{*b}, \ldots, X_n^{*b}$ is artificially simulated by resampling with replacement. In other words, after the extraction of an element, this is replaced in the original sample such that it can be chosen again.
2. For each sample obtained, the value of the statistic $T^{*b} = T(X_1^{*b}, \ldots, X_n^{*b})$ is calculated.
3. Steps 1 and 2 are repeated a large number (B) of times so as to obtain bootstrap values $T^{*1}, \ldots, T^{*B}$. At this stage in its development, computer software allows for the computational cost entailed in the generation of a large number of samples to be estimated.
4. Finally, the distribution of T and its corresponding quantiles is approximated by means of a histogram obtained on the basis of values $T^{*1}, \ldots, T^{*B}$, that is, an empirical approach to the sample distribution of statistic T is obtained, without any assumptions having been made as to the theoretical distribution to which the latter conforms.

The bootstrap resampling method outlined above is known as naive bootstrap. In step 1, the bootstrap samples are simulated by means of resampling with replacement, that is, based on the empirical distribution $\hat{F_n}(x) = n^{-1} \sum_{i=1}^{n} I_{\{X_i \leq x\}}$ of the sample. Other resampling approaches have also been considered: When the form of the population distribution is known, the use of parametric bootstrap allows for a better approximation of the sampling distribution of the test statistic. That is to say, if F is known to belong to a parametric family $\{F_\theta : \theta \in \Theta\}$ and $\hat{\theta}$ is an estimator of θ (e.g., the estimator of maximum likelihood), then $\mathbf{X}^*$ can be taken as a random sample from $F_{\hat{\theta}}$. For instance, assuming one knows that the observations came from a normal distribution $X \in N(\mu, \sigma)$, then one would draw repeated bootstrap samples from a normal distribution $N(\bar{X}, S)$. Where the variable X is assumed to be continuous with the density function, f, the use of smoothed bootstrap might be more appropriate. Instead of resampling directly from the empirical distribution $\hat{F_n}(x)$, one first smoothes it out. This smoothed version is then used to generate new samples from the smoothed distribution $\tilde{F_n}(x) = \int_{-\infty}^{x} \hat{f_n}(x)\, \mathrm{d}x$, where $\hat{f_n}(x)$ is an estimate of density function $f(x)$.

As previously pointed out, rather than being perceived as an isolated method, bootstrap is instead used in combination with other statistical techniques. As a review of all bootstrap applications published in recent years would prove extremely difficult, the main thrust of this chapter will focus on how bootstrap works in a regression context. In particular, the possibilities of bootstrap will be analyzed in the context of generalized additive models (GAMs). These types of models were proposed by Hastie and Tibshirani[7] as a unifying family of flexible models covering a wide range of regression models with different types of responses (e.g., Gaussian, binomial, and Poisson).

Rather than seeking to provide a comprehensive overview of all bootstrap applications in this regression context, this chapter will confine itself to discussing bootstrap resampling methods for achieving two different goals in a GAM context, namely (1) construction of CIs for covariate effects on response and (2) implementation of tests capable of detecting the statistical significance of interaction among the effects of the different covariates.

This chapter is laid out as follows: Section 1.06.2 presents different bootstrap resampling methods for regression; Section 1.06.3 introduces GAMs and briefly discusses the process of estimating these models; and

Section 1.06.4 addresses the construction of bootstrap-based CIs for covariate effects in GAMs. The advantage of bootstrap methodology versus a classical method of constructing asymptotic intervals is illustrated in Section 1.06.4.1, by reference to air pollution time-series studies.[8]

The effect of a given covariate on response may often vary with the values taken by another covariate, something that in turn leads to the concept of interaction. Section 1.06.5 introduces GAMs that include interactions among covariate effects. Section 1.06.6 poses the problem of test hypotheses for the detection of significant interactions in GAMs. Finally, Section 1.06.6.1 contains an application to real SO_2 binary pollution time-series data.[9]

1.06.2 Bootstrap Resampling Methods for Regression

In many fields of research, it is important to establish the relationship between a response variable of interest and one or more explanatory covariates. Regression studies enable mathematical models to be obtained that link the conditional mean $\mu(\mathbf{X}) = E(Y|\mathbf{X})$ of response Y to p covariates $\mathbf{X} = \left(X_1, \ldots, X_p\right)$. This section sets out different bootstrap resampling procedures for regression data sets $\left(\mathbf{X}_{1,} Y_1\right), \ldots, (\mathbf{X}_{\mathrm{n}}, Y_n)$ of $(\mathbf{X}, Y)$. These procedures will be used in Section 1.06.4 to construct CIs for covariate effects, and in Section 1.06.6 to test the significance of interaction terms.

Let us assume that the relationship between covariate $\mathbf{X}$ and response variable Y can be expressed as $Y = \mu(\mathbf{X}) + \varepsilon$, where ε is an error variable having zero mean $E(\varepsilon|\mathbf{X}) = 0$. Bootstrap samples $\left(\mathbf{X}_1^*, Y_1^*\right), \ldots, \left(\mathbf{X}_n^*, Y_n^*\right)$ are generated by being drawn from some estimated distribution $\hat{R_n}$ of the true distribution of $(\mathbf{X}, Y)$. If $\hat{R_n}$ is chosen as the empirical distribution of $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, then the bootstrap samples are generated by being drawn with replacement from the original sample. This bootstrap procedure is called pairwise resampling.[10,11] Another alternative is the use of smooth bootstrap described above, where the bootstrap samples are drawn from a smoothed distribution estimate of the joint density of $\mathbf{X}$ and Y. For a more detailed discussion of this bootstrap, the reader should refer to Cao-Abad and González-Manteiga.[12]

In bootstrap procedures proposed until now, both $\mathbf{X}$ and Y are assumed to be random variables. This sample scheme is known as random design. However, there are applications where the researcher is able to control the values of the covariates $\mathbf{X}$, and Y is the only random variable. The scheme known as fixed design will now be discussed. The first point to be borne in mind here is that the error variables ε_i are independently and identically distributed (i.i.d.) with zero mean. In this case, the bootstrap samples $\left(\mathbf{X}_1, Y_1^*\right), \ldots, \left(\mathbf{X}_n, Y_n^*\right)$ can be obtained with $Y_i^* = \tilde{\mu}(\mathbf{X}_i) + \hat{\varepsilon}_i^*$, where $\tilde{\mu}(\mathbf{X}_i)$ is a pilot estimate of $\mu(\mathbf{X}_i)$ and $\hat{\varepsilon}_1{}^*, \ldots, \hat{\varepsilon}_n^*$ is a random sample drawn from the empirical distribution of the centered residuals $\hat{\varepsilon}_i = Y_i - \tilde{\mu}(\mathbf{X}_i)$. This bootstrap procedure is called residual resampling. Errors cannot be assumed to be i.i.d. in all cases, for example, this condition is not fulfilled when the response variance depends on covariate $\mathbf{X}$. In a situation of this kind, possibly the most popular resampling method is the so-called wild bootstrap[13–15] introduced by Wu.[16] It allows for heterogeneous variance in the residuals. In wild bootstrap, each residual $\hat{\varepsilon}_i^*$ is drawn from a distribution that seeks to imitate the distribution of $\hat{\varepsilon}_i$ up to the first three moments, namely, $E[\hat{\varepsilon}_i^*] = 0$, $E[\hat{\varepsilon}_i^{*2}] = \hat{\varepsilon}_i{}^2$, and $E[\hat{\varepsilon}_i^{*3}] = \hat{\varepsilon}_i^3$. These conditions are met if ε_i^* is drawn from the two-point distribution, with probability mass at $a = 0.5\left(1 - \sqrt{5}\right)\hat{\varepsilon}_i$ and $b = 0.5\left(1 + \sqrt{5}\right)\hat{\varepsilon}_i$ occurring with probabilities $q = 0.1\left(5 + \sqrt{5}\right)$ and $1 - q$, respectively. A fuller review of resampling methods in regression can be found in Schimek.[17]

The resampling methods proposed thus far have basically been designed for continuous response regression models. Yet this type of resampling does not adapt to other types of distribution, for example, for binary response models, values other than zero or one, Y_i^*, will be obtained, with the result that the model will perform in quantitative but not qualitative terms. Similarly, should there be a Poisson response, there is no guarantee that the values Y_i^* will not be negative. A general way of approaching these types of situations is to consider the exponential family,[18] which includes the majority of distributions in statistics, that is, Gaussian, binomial, or Poisson. In the exponential family, the conditional density function $f(Y|\mathbf{X})$ is given by

$$f(Y|\mathbf{X}) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \tag{1}$$

where θ denotes the natural parameter, ϕ is the scale or dispersion parameter, and a, b, and c are known specific functions in the exponential family. The natural parameter θ depends on the conditional mean of the response $\mu(X) = E(Y|\mathbf{X})$, and so one can write $\theta = \theta(\mu(X))$. Moreover, the conditional variance of the response, $\mathrm{Var}(Y|\mathbf{X})$, depends on $\mu(X)$, via the relationship $\mathrm{Var}(Y|\mathbf{X}) = \hat{\phi} V(\mu(\mathbf{X}))$, where V is the so-called variance function, which is known and is determined by the exponential distribution chosen.

Bootstrap resampling techniques for obtaining bootstrap samples are based on the conditional distribution of $Y|\mathbf{X}$, determined by $\mu(\mathbf{X})$ and ϕ, with the result that the density function $f(Y|\mathbf{X})$ given in Equation (1) can therefore be written as $f(\mu(\mathbf{X}), \phi)$. The bootstrap technique consists of generating a large number (B) of bootstrap replicates $\left(\mathbf{X}_{1,} Y_1^{*b}\right), \ldots, \left(\mathbf{X}_n, Y_n^{*b}\right)$ ($b = 1, \ldots, B$), with $Y_i^{*b} \sim f(\tilde{\mu}(\mathbf{X}_i)\phi)$ being based on the original sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, and $\tilde{\mu}(\mathbf{X}_i)$ being the pilot estimates obtained on this basis.[19,20] In special situations, the dispersion parameter ϕ is known and the conditional distribution of Y is specified by $\mu(\mathbf{X})$. For instance, for binary response $Y_i \sim \mathrm{Bernoulli}(\mu(\mathbf{X}_i))$, the bootstrap replicates will be generated in line with $Y_i^* \sim \mathrm{Bernoulli}(\tilde{\mu}(\mathbf{X}_i))$. Likewise, for Poisson response $Y_i \sim \mathrm{Poisson}(\mu(\mathbf{X}_i))$, the bootstrap replicates will be generated in line with $Y_i^* \sim \mathrm{Poisson}(\tilde{\mu}(\mathbf{X}_i))$. Otherwise, if ϕ is unknown, an estimate $\hat{\phi}$ for parameter ϕ can be obtained, in line with $\mathrm{Var}(Y|\mathbf{X}) = \hat{\phi} V(\mu(\mathbf{X}))\mathrm{Var}(Y|\mathbf{X}) = \hat{\phi} V(\mu(\mathbf{X}))$ or equivalently $\phi = \mathrm{Var}(Y|\mathbf{X})/V(\mu(\mathbf{X}))$, from

$$\hat{\phi} = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}\,(\mathbf{X}_i))^2}{nV(\hat{\mu}(\mathbf{X}_i))}$$

where $\hat{\mu}(\mathbf{X}_i)$ are the estimates of $\mu(\mathbf{X}_i)$ obtained on the basis of the original sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n}$. For instance, in the case of Gaussian response $Y_i - N(\mu(X_i), \sigma)$ the parameter ϕ is unknown and coincides with σ^2. In such a case, the bootstrap responses will be generated in line with $Y_i^* - N(\tilde{\mu}(X_i), \hat{\sigma})$, where $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \tilde{\mu}\,(X_i))^2$ is the estimation of σ^2. It should be noted that in this approach, the variance σ^2 is a constant and the regression model is thus assumed to be homoscedastic. Nevertheless, our interest also lies in heteroscedastic regression models, where the variance $\sigma^2 = \sigma^2(\mathbf{X})$ depends on the vector of the covariates $\mathbf{X}$. To estimate $\sigma^2(\mathbf{X})$, a nonparametric regression model can be fitted to $R_i = (Y_i - \tilde{\mu}\,(\mathbf{X}_i))^2$ on $\mathbf{X}$. Another alternative is to use the wild bootstrap method outlined above.

1.06.3 Generalized Additive Models

The models that will be studied here can be viewed as a generalization of the well-known generalized linear model (GLM).[18] In these GLMs, the conditional mean $\mu(\mathbf{X}) = E[Y|\mathbf{X}]$ of response Y depends on the covariates $\mathbf{X} = \left(X_1, \ldots, X_p\right)$ via $\mu(\mathbf{X}) = H\left(\alpha + \alpha_1 X_1 + \cdots + \alpha_p X_p\right)$, where H is a known monotone link function and $\left(\alpha, \alpha_1, \ldots, \alpha_p\right)$ is a vector of coefficients. In some instances, GLMs can be very restrictive, because they assume linearity in the covariates. This constraint can be avoided by replacing the linear index $\eta = \alpha + \alpha_1 X_1 + \cdots + \alpha_p X_p$ with a nonparametric structure. Accordingly, here we will concentrate on the GAM[7], which is a generalization of the GLM, by introducing one-dimensional, nonparametric functions in lieu of linear components. Specifically, GAMs express the conditional mean $\mu(\mathbf{X})$ as

$$\mu(\mathbf{X}) = H\left(\alpha + f_1(X_1) + \cdots + f_p(X_p)\right) \tag{2}$$

where H is a known link function, α is a constant, and $f_1, \ldots, f_p$ are unspecified, unknown, zero-mean functions. In the case of an identity function, H, one speaks of an additive model. On assuming that effects are additive, GAMs maintain the interpretability of GLMs.[18,21] Yet, at the same time, they incorporate the flexibility of nonparametric smoothing methods because, rather than following a fixed parametric form, the effect of each of the covariates, X_j, depends on a totally unknown function, f_j, which is only required to possess a certain degree