# Predictive distributions for non-regular parametric models

Giovanni Fonseca[1], Federica Giummolè[2], Paolo Vidoni[1]

[1] University of Udine, Department of Economics and Statistics, via Treppo 18, I-33100 Udine, ITALY. e-mail: giovanni.fonseca@uniud.it, paolo.vidoni@uniud.it

[2] Ca' Foscari University - Venice, Department of Environmental Sciences, Informatics and Statistics, San Giobbe, Cannaregio 783, I-30121 Venice, ITALY. e-mail: giummole@unive.it

**Abstract:** Improved prediction distributions based on asymptotic methods are a well known tool for prediction in the context of regular parametric models. On the contrary, for non-regular cases, prediction is mainly based on the estimative or plug-in distribution. The aim of this work is to define calibrated predictive distributions which quantiles have coverage probability equal or close to the target nominal value. Whenever the computation is not feasible, a suitable bootstrap procedure easily provides a good estimate for the proposed distribution. A simulation example is provided for a particular non regular model, the generalized extreme value distribution, which support depends on unknown parameters.

**Keywords:** Coverage probability; Extreme value distributions; Non-regular models; Parametric bootstrap; Prediction limits; Predictive distributions.

## 1 Introduction

In this work, we consider the problem of prediction of a future, or unobservable, unidimensional absolutely continuous random variable $Z$, on the basis of an observed sample $y = (y_1, \ldots, y_n)$ from a random vector $Y = (Y_1, \ldots, Y_n)$. We assume that the joint distribution of $(Y, Z)$ is known, up to a $k$-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^k$. In this case, a possible solution can be given in terms of prediction limits, i.e. functions $\tilde{z}_\alpha(\hat{\theta})$ such that, for all $\alpha \in (0, 1)$, the coverage probability

$$P_{Y,Z}\left[Z \leq \tilde{z}_\alpha(\hat{\theta}(Y))\right] = \alpha, \tag{1}$$

at least to a high order of approximation. Here $\hat{\theta} = \hat{\theta}(Y)$ is an asymptotically efficient estimator for $\theta$, usually the maximum likelihood estimator. When exact results are not available, an easy solution is given by considering the estimative prediction limits, obtained by substituting the unknown parameter $\theta$ by $\hat{\theta}$ in the $\alpha$-quantiles of the conditional distribution of $Z$ given $Y = y$. Unfortunately the associated coverage error has order $O(n^{-1})$, which is often considerable. Improved prediction limits with coverage error of order $o(n^{-1})$ have been proposed by Barndorff-Nielsen and

Cox(1996) and Vidoni (1998), as modifications of the estimative prediction limits. Their results rely on asymptotic expansions and only hold under regularity assumptions on the model. Calibrated prediction limits can be obtained by means of a bootstrap based procedure, as proposed by Hall et al. (1999). Though very interesting, this approach provides solutions for specific fixed values of the target coverage $\alpha$.
In this work, following Fonseca et al. (2010), we define a predictive distribution which $\alpha$-quantiles provide exact prediction limits for every $\alpha \in (0, 1)$. When this predictive distribution is not explicitly available, it can be approximated using a suitable bootstrap technique. The coverage error associated to the resulting approximated quantiles is of order $o(n^{-1})$, improving on the estimative solution. The proposed method for prediction is general, easy to compute and does not require regularity assumptions on the underlying model. Thus, it also applies to non-regular cases when the support of the model depends on an unknown parameter. This extension is very useful, for instance, in the applications to survival analysis and in the studies of extreme events.

## 2 Calibrated predictive distributions

Let us assume, for simplicity, that $Y_1, \ldots, Y_n, Z$ are independent continuous random variables with the same distribution. Denote by $G(z; \theta)$ the distribution function of $Z$.
Consider the estimative prediction limit $z_\alpha(\hat{\theta}) = G^{-1}(\alpha; \hat{\theta})$, where $G^{-1}(\cdot; \hat{\theta})$ is the inverse of function $G(\cdot; \hat{\theta})$. The associated coverage probability is

$$P_{Y,Z}\{Z \leq z_\alpha(\hat{\theta}); \theta\} = E_Y[G\{z_\alpha(\hat{\theta}); \theta\}; \theta] = C(\alpha, \theta).$$

Function $C(\alpha, \theta)$ depends on the true parameter value $\theta$ and on the nominal coverage probability $\alpha$. However, its explicit expression is rarely available. It is well known that it does not match the target value $\alpha$, although asymptotically $C(\alpha, \theta) = \alpha + O(n^{-1})$, as $n \to +\infty$.
As suggested by Fonseca et al. (2010), a predictive distribution function can be defined by substituting $\alpha$ with $G(z; \hat{\theta})$ in $C(\alpha, \theta)$:

$$G_c(z; \hat{\theta}, \theta) = C\{G(z; \hat{\theta}), \theta\}. \tag{2}$$

$G_c(\cdot; \hat{\theta}, \theta)$ is a proper predictive distribution function in regular parametric models. When the support of $Z$ depends on $\theta$, $G_c(z; \hat{\theta}, \theta)$ may not satisfy one or both the limit conditions as $z \to \infty$. Nevertheless, it can still be fruitfully employed for obtaining good prediction limits, far from the boundary of the support of $Z$.
The predictive distribution (2) gives, as quantiles, prediction limits $z^c_\alpha(\hat{\theta}, \theta)$ which coverage probability equals the target nominal value $\alpha$, for all $\alpha \in (0, 1)$.

Though interesting from a theoretical perspective, the calibrated predictive distribution $G_c(z;\hat{\theta},\theta)$ is in fact inapplicable since it usually depends on the unknown parameter $\theta$. A useful surrogate is the corresponding plug-in estimator

$$\hat{G}_c(z;\hat{\theta}) = G_c(z;\hat{\theta},\hat{\theta}) = C\{G(z;\hat{\theta}),\hat{\theta}\}.$$

The associated $\alpha$-prediction limit is defined as $\hat{z}^c_\alpha(\hat{\theta}) = z^c_\alpha(\hat{\theta},\hat{\theta}) = z_{\hat{\alpha}^c}(\hat{\theta})$, with $\hat{\alpha}_c = C^{-1}(\alpha,\hat{\theta})$, and it satisfies (1) to a closer approximation than the estimative prediction limit $z_\alpha(\hat{\theta})$, that is with an error term of order $o(n^{-1})$.

A closed form expression for the coverage probability $C(\alpha,\theta)$ is rarely available so that even the predictive distribution function $\hat{G}_c(z;\hat{\theta})$ is not very useful in practice. Anyway, there is a suitable parametric bootstrap estimator for $G_c(z;\hat{\theta},\theta)$, to be considered when $C(\alpha,\theta)$ is not available. Let $y^*(j)$, $j = 1,\ldots,B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^*(j)$, $j = 1,\ldots,B$, be the corresponding maximum likelihood estimates. Since $C(\alpha,\theta) = E_Y[G\{z_\alpha(\hat{\theta});\theta\};\theta]$, we define the bootstrap-calibrated predictive distribution as

$$G^b_c(z;\hat{\theta}) = \frac{1}{B}\sum_{j=1}^{B} G\{z_\alpha(\hat{\theta}^*_j);\hat{\theta}\}|_{\alpha=G(z;\hat{\theta})}. \tag{3}$$

The corresponding $\alpha$-quantile defines, for each $\alpha \in (0,1)$, a prediction limit having coverage probability equal to the target $\alpha$, with an error term which depends on the efficiency of the bootstrap simulation procedure. It is important noticing that the computation of (3) does not require any assumption on the regularity of the parametric models involved, as long as the bootstrap applies.

## 3 Generalized extreme value distribution

Let $Y_1,\ldots,Y_n$ be independent random variables with common generalized extreme value distribution, that is

$$G(y;\mu,\sigma,\xi) = \exp\left\{-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right\},$$

where $1+\xi(y-\mu)/\sigma > 0$ and $\theta = (\mu,\sigma,\xi)$ is an unknown parameter with $\sigma > 0$ a scale parameter, $\mu \in \mathbb{R}$ a location parameter and $\xi \in \mathbb{R}$ a shape parameter. The generalized extreme value distribution includes the Frechet, the Gumbel and the Weibull distributions as particular cases and is usually used for the study of extreme events, such as extreme flood of a river or maximum sea level. In this context it can be useful to consider the problem of prediction of a future value $Z = Y_{n+1}$, independent of $Y_1,\ldots,Y_n$ and with the same distribution.

TABLE 1. Generalized extreme value distribution. Coverage probabilities for estimative and bootstrap calibrated prediction limits of level $\alpha$=0.9, 0.95, 0.99.

| $\alpha$ | $n$ | Estimative | Bootstrap |
|---|---|---|---|
| 0.9 | 10 | 0.880 | 0.899 |
| | 20 | 0.893 | 0.905 |
| 0.95 | 10 | 0.933 | 0.954 |
| | 20 | 0.942 | 0.951 |
| 0.99 | 10 | 0.976 | 0.987 |
| | 20 | 0.982 | 0.986 |

In this case, an explicit expression for the coverage probability $C(\alpha, \mu, \sigma, \xi)$, associated to the estimative $\alpha$-prediction limit, is not available. As explained in Section 2, we can estimate (2) using the bootstrap estimator (3) and calculate calibrated prediction limits as quantiles of this approximated predictive distribution.
Table 1 shows the results of a simulation study for comparing the performance of estimative (Estimative) and bootstrap calibrated (Bootstrap) prediction limits, with respect to the corresponding coverage probabilities. Estimation is based on 5,000 Monte Carlo replications. Bootstrap procedure is based on 1,000 bootstrap samples. Estimated standard errors are always smaller than 0.005. Different values of the target level, $\alpha$=0.9, 0.95, 0.99, and of the sample size, $n = 10, 20$, are considered. The parameters of the generalized extreme value model are fixed to $\mu = 5$, $\sigma = 2$ and $\xi = 0.4$. It can be seen that the bootstrap solution remarkably improves on the estimative one.

## References

Barndorff-Nielsen, O.E., and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli*, **2**, pp. 319-340.

Fonseca, G., Giummolè, F., and Vidoni, P. (2010). Calibrating predictive distributions. *Redazioni Provvisorie*, **2/2010**, Department of Statistics, Ca' Foscari University, Venice.

Hall, P., Peng, L. and Tajvidi, N. (1999). On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika*, **86**, pp. 871-880.

Vidoni, P. (1998). A note on modified estimative prediction limits and distributions. *Biometrika*, **85**, pp. 949-953.