Model-free inference in statistics: how and why.

Dimitris N. Politis
Department of Mathematics
University of California at San Diego
La Jolla, CA 92093-0112;
email: dpolitis@ucsd.edu

1 Estimation

Parametric models served as the cornerstone for the foundation of Statistical Science in the beginning of the 20th century by R.A. Fisher, K. Pearson, J. Neyman, E.S. Pearson, W.S. Gosset (also known as "Student"), etc.; their seminal developments resulted into a complete theory of statistics that could be practically implemented using the technology of the time, i.e., pen and paper (and slide-rule!). While some models are inescapable, e.g. modeling a polling dataset as a sequence of independent Bernoulli random variables, others appear contrived, often invoked for the sole reason to make the mathematics work. As a prime example, the ubiquitous—and typically unjustified—assumption of Gaussian data permeates statistics textbooks to the day. Model criticism and diagnostics were subsequently developed as a practical way out.

With the advent of widely accessible powerful computing in the late 1970s, computer-intensive methods such as resampling and cross-validation created a revolution in modern statistics. Using computers, statisticians became able to analyze big datasets for the first time, paving the way towards the 'big data' era of the 21st century. But perhaps more important was the realization that the way we do the analysis could/should be changed as well, as practitioners were gradually freed from the limitations of parametric models. For instance, the great success of Efron's (1979) bootstrap was in providing a complete theory for statistical inference under a nonparametric setting much like Maximum Likelihood Estimation had done half a century earlier under the restrictive parametric setup.

Nevertheless, there is a further step one may take, i.e., going beyond even nonparametric models. To explain this, let us first focus on regression, i.e., data that are pairs: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ where Y_i is the measured response associated with a regressor value of X_i. The standard homoscedastic additive model in this situation reads:

$$Y_i = \mu(X_i) + \epsilon_i \tag{1}$$

where the random variables ϵ_i are assumed to be independent, identically distributed (i.i.d.) from a distribution $F(\cdot)$ with mean zero.

- **Parametric model:** Both $\mu(\cdot)$ and $F(\cdot)$ belong to parametric families of functions, i.e., a setup where the only unknown is a finite-dimensional parameter; a typical example is straight-line regression with Gaussian errors, i.e., $\mu(x) = \beta_0 + \beta_1 x$ and $F(\cdot)$ being $N(0, \sigma^2)$.
- **Semiparametric model:** $\mu(\cdot)$ belongs to a parametric family, whereas $F(\cdot)$ does not; instead, it may be assumed that $F(\cdot)$ belongs to a smoothness class, e.g., assume that $F(\cdot)$ is absolutely continuous.
- **Nonparametric model:** Neither $\mu(\cdot)$ nor $F(\cdot)$ can be assumed to belong to parametric families of functions.

Despite the nonparametric aspect of it, even the last option constitutes a model, and can thus be rather restrictive. To see why, note that eq. (1) with i.i.d. errors is not satisfied in many cases of interest even after allowing for heteroscedasticity of the errors. Nevertheless, it is possible to shun eq. (1) altogether and instead adopt a *model-free* setup that can be described as follows.

- **Model-free regression:**
  - **Random design.** The pairs $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ are i.i.d.
  - **Deterministic design.** The variables $X_1, \ldots, X_n$ are deterministic, and the random variables $Y_1, \ldots, Y_n$ are independent with common conditional distribution, i.e., $P\{Y_j \leq y | X_j = x\} = D_x(y)$ not depending on j.

Inference for features, i.e. functionals, of the common conditional distribution $D_x(\cdot)$ is still possible under some regularity conditions, e.g. smoothness. Arguably, the most important such feature is the conditional mean $E(Y|X = x)$ that can be denoted $\mu(x)$. When $\mu(x)$ can be assumed smooth, it can be consistently estimated by a local average and/or local polynomial. Asymptotic normality and/or resampling can then be invoked to construct confidence intervals for $\mu(x)$.

2 Prediction

Traditionally, the problem of prediction has been approached in a model-based way, i.e., (a) fit a model such as (1), and then use the fitted model for prediction of a future response Y_f associated with a regressor value x_f. Note that even in the absence of model (1), the conditional expectation $\mu(x_\mathrm{f})$=$E(Y_\mathrm{f}|X_\mathrm{f} = x_\mathrm{f})$ is the Mean Squared Error (MSE) optimal predictor of Y_f. As already mentioned, $\mu(x_\mathrm{f})$ can be estimated in a model-free way and then used for predicting Y_f but a problem remains: how to gauge the accuracy of prediction, i.e., how to construct a prediction—as opposed to confidence—interval.

Interestingly, it is possible to accomplish the goal of point and interval prediction of Y_f under the model-free regression setup in a direct fashion, i.e., without the intermediate step of model-fitting; this is achieved via the **Model-free Prediction Principle** expounded upon in Politis (2015). Model-Free Prediction restores the emphasis on observable quantities, i.e., current and future data, as opposed to unobservable model parameters and estimates thereof. In this sense, the Model-Free Prediction Principle is in concordance with Bruno de Finetti's statistical philosophy. Notably, being able to predict the response Y_f associated with the regressor X_f taking on *any* possible value (say x_f) seems to inadvertently also achieve the main goal of modeling, i.e., trying to relate how Y depends on X. In so doing, the solution to interesting estimation problems is obtained as a by-product, e.g. inference on features of $D_x(\cdot)$ such as its mean $\mu(x)$. In other words, as prediction can be treated as a by-product of model-fitting, key estimation problems can be solved as a by-product of being able to perform prediction. Hence, a Model-free approach to frequentist statistical inference is possible, including prediction and confidence intervals.

3 The Model-free Prediction Principle

Consider the model-free regression set-up with a vector of observed responses $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ that are associated with the vector of regressors $\underline{X}_n = (X_1, \ldots, X_n)'$. Also consider the enlarged vectors $\underline{Y}_{n+1} = (Y_1, \ldots, Y_n, Y_{n+1})'$ and $\underline{X}_{n+1} = (X_1, \ldots, X_n, X_{n+1})'$ where (Y_{n+1}, X_{n+1}) is an alternative notation for $(Y_\mathrm{f}, X_\mathrm{f})$; recall that Y_f is yet unobserved, and X_f will be set equal to the value x_f of interest. If the Y_is were i.i.d. (and not depending on their associated X value), then prediction would be trivial: the MSE–optimal predictor of Y_{n+1} is simply given by the common expected value of the Y_is, completely disregarding the value of X_{n+1}.

In a nutshell, the Model-Free Prediction Principle amounts to using the structure of the problem in order to **find an invertible transformation H_m that can map the non-i.i.d. vector $\underline{Y}_m$ to a vector $\underline{\epsilon}_m = (\epsilon_1, \ldots, \epsilon_m)'$ that has i.i.d. components;** here m could be taken equal to either n or $n+1$ as needed. Letting H_m^{-1} denote the inverse transformation, we have $\underline{\epsilon}_m = H_m(\underline{Y}_m)$ and $\underline{Y}_m = H_m^{-1}(\underline{\epsilon}_m)$, i.e.,

$$\underline{Y}_m \stackrel{H_m}{\longmapsto} \underline{\epsilon}_m \quad \text{and} \quad \underline{\epsilon}_m \stackrel{H_m^{-1}}{\longmapsto} \underline{Y}_m. \tag{2}$$

If the practitioner is successful in implementing the Model-Free procedure, i.e., in identifying (and estimating) the transformation H_m to be used, then the prediction problem is reduced to the trivial one of predicting i.i.d. variables. To see why, note that eq. (2) with $m = n+1$ yields $\underline{Y}_{n+1} = H_{n+1}^{-1}(\underline{\epsilon}_{n+1}) = H_{n+1}^{-1}(\underline{\epsilon}_n, \epsilon_{n+1})$. But $\underline{\epsilon}_n$ can be treated as known (and constant) given the data $\underline{Y}_n$; just use eq. (2) with $m = n$. Since the unobserved Y_{n+1} is just the $(n+1)^{th}$ coordinate of vector $\underline{Y}_{n+1}$, we have just expressed Y_{n+1} as a function of the unobserved ϵ_{n+1}. Note that predicting a function, say $g(\cdot)$, of an i.i.d. sequence $\epsilon_1, \ldots, \epsilon_n, \epsilon_{n+1}$ is straightforward because $g(\epsilon_1), \ldots, g(\epsilon_n), g(\epsilon_{n+1})$ is simply another sequence of i.i.d. random variables. Hence, the practitioner can use this simple structure to develop point predictors for the future response Y_{n+1}.

Prediction intervals can then be immediately constructed by resampling the i.i.d. variables $\epsilon_1, \ldots, \epsilon_n$; this can be thought to give an extension of the model-based, residual bootstrap of Efron (1979) to model-free settings since, if model (1) were to hold true, the residuals from the model could be considered as the outcomes of the requisite transformation H_n.

4 Time series

Under regularity conditions, a transformation such as H_m of the Model-Free Prediction Principle always exists but is not necessarily unique. For example, if the variables $(Y_1, \ldots, Y_m)$ have an absolutely continuous joint distribution and no explanatory variables $\mathbf{X}_m$ are available, then the Rosenblatt (1952) transformation can map them onto a set of i.i.d. random variables. Nevertheless, estimating the Rosenblatt transformation from data may be infeasible except in special cases. On the other hand, a practitioner may exploit a given structure for the data at hand, e.g., a regression structure, in order to construct a different, case-specific transformation that may be practically estimable from the data.

Recall that the Rosenblatt transformation maps an arbitrary random vector $\underline{Y}_m = (Y_1, \ldots, Y_m)'$ having absolutely continuous joint distribution onto a random vector $\underline{U}_m = (U_1, \ldots, U_m)'$ whose entries are i.i.d. Uniform(0,1); this is done via the probability integral transform based on conditional distributions. For $k > 1$ define the conditional distributions $F_k(y_k|y_{k-1}, \ldots, y_1) = P\{Y_k \leq y_k | Y_{k-1} = y_{k-1}, \ldots, Y_1 = y_1\}$, and let $F_1(y_1) = P\{Y_1 \leq y_1\}$. Then the Rosenblatt transformation amounts to letting

$$\begin{aligned} U_1 = F_1(Y_1), U_2 = F_2(Y_2|Y_1), U_3 = F_3(Y_3|Y_2, Y_1), \\ \ldots, \text{ and } U_m = F_m(Y_m|Y_{m-1}, \ldots, Y_2, Y_1). \end{aligned} \tag{3}$$

The problem is that the distributions F_k for $k \geq 1$ are typically unknown and must be estimated (in a continuous fashion) from the $\underline{Y}_n$ data at hand. However, unless there is some additional structure, this estimation task may be unreliable or plain infeasible for large k. As an extreme example, note that to estimate F_n we would have only one point (in n-dimensional space) to work with. Hence, without additional assumptions, the estimate of F_n would be a point mass which is a completely unreliable estimate, and of little use in terms of constructing a probability integral transform due to its discontinuity.

An example of additional structure is the Markov setup. To elaborate, suppose that the data $Y_1, \ldots, Y_n$ are a realization of a stationary (and ergodic) Markov chain. In this case, the conditional distributions F_k for all $k > 1$ are completely determined by the one-step transition distribution, namely F_2. To see why, note that the Markov assumption implies that $P\{Y_k \leq y_k | Y_{k-1} = y_{k-1}, \ldots, Y_1 =$

$y_1\} = P\{Y_k \leq y_k | Y_{k-1} = y_{k-1}\}$ for $k > 1$. Hence, the practitioner may use kernel smoothing or a related technique on the data pairs $\{(Y_j, Y_{j+1})$ for $j = 1, \ldots, n-1\}$ in order to estimate the common joint distribution of these pairs. In turn, this yields estimates of F_1 and F_2, and by extension F_k for $k > 2$, so that the Rosenblatt transformation can be practically implemented as part of the Model-Free Prediction Principle. Further examples of transformations applicable to diverse settings with regression and/or time series data are discussed in Politis (2015).

References

[1] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.*, vol. 7, pp. 1–26.

[2] Politis, D.N. (2015). *Model–free Prediction and Regression: a Transformation-based Approach to Inference*, Springer, New York.

[3] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.*, vol. 23, pp. 470–472.