Model-free model-fitting and predictive distributions

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

Abstract

The problem of prediction is revisited with a view towards going beyond the typical nonparametric setting and reaching a fully model-free environment for predictive inference, i.e., point predictors and predictive intervals. A basic principle of model-free prediction is laid out based on the notion of transforming a given setup into one that is easier to work with, namely i.i.d. or Gaussian. As an application, the problem of nonparametric regression is addressed in detail; the model-free predictors are worked out, and shown to be applicable under minimal assumptions. Interestingly, model-free prediction in regression is a totally automatic technique that does not necessitate the search for an optimal data transformation before model fitting. The resulting model-free predictive distributions and intervals are compared to their corresponding model-based analogs, and the use of cross-validation is extensively discussed. As an aside, improved prediction intervals in linear regression are also obtained.

Keywords: Bootstrap, cross-validation, frequentist prediction, heteroskedasticity, nonparametric estimation, prediction intervals, regression, smoothing, transformations.

Acknowledgement. A preliminary version of this paper was presented as a Plenary Talk at the 10th International Vilnius Conference on Probability and Mathematical Statistics, June 28–July 3, 2010, and as a Special Invited Talk at the 28th European Meeting of Statisticians, August 17-22, 2010; the author is grateful to the audiences in these two—and several other—occasions for their helpful feedback. Many thanks are due to Arthur Berg, Wilson Cheung and Tim McMurry for invaluable help with R functions and computing, and to Richard Davis, Jeff Racine, Bill Schucany, Dimitrios Thomakos and Slava Vasiliev for helpful discussions. The author is also grateful to the Editors, Ricardo Cao and Domingo Morales, for their support and encouragement, and to six (!) anonymous referees for their very detailed and constructive comments; one of the referees deserves special thanks for an astute observation that helped shed light on the workings of the 'uniformize'

algorithm of Section 4. This work has been partially supported by NSF grants DMS-07-06732 and DMS-10-07513, and by a fellowship from the Guggenheim Foundation.

1 Introduction

In the classical setting of an i.i.d. (independent and identically distributed) sample, the problem of prediction is not very interesting. Consequently, practitioners have mostly focused on estimation and hypothesis testing in this case. However, when the i.i.d. assumption no longer holds, the prediction problem is both important and intriguing; see Geisser (1993) for an introduction. Typical examples where the i.i.d. assumption breaks down include regression problems and dependent data.

Two key models are given below.

- **Regression**

$$Y_t = \mu(\underline{x}_t) + \sigma(\underline{x}_t)\ \varepsilon_t \quad \text{for } t = 1, \ldots, n. \tag{1}$$

- **Time series**

$$Y_t = \mu(Y_{t-1}, \cdots, Y_{t-p}; \underline{x}_t) + \sigma(Y_{t-1}, \cdots, Y_{t-p}; \underline{x}_t)\ \varepsilon_t \quad \text{for } t = 1, \ldots, n. \tag{2}$$

Here, $Y_1, \ldots, Y_n$ are the data, ε_t are the errors assumed i.i.d. $(0, 1)$, and $\underline{x}_t$ is a fixed-length vector of explanatory (predictor) variables associated with the observation Y_t. The functions $\mu(\cdot)$ and $\sigma(\cdot)$ are unknown but assumed to belong to a class of functions that is either finite-dimensional (parametric family) or not; the latter case is the usual nonparametric setup in which case the functions $\mu(\cdot)$ and $\sigma(\cdot)$ are typically assumed to belong to a smoothness class.

Given one of these two models, the optimal *model-based* predictors of a future Y-value can be constructed. Nevertheless, the prediction problem can, in principle, be carried out in a fully model-free setting, offering—at the very least—robustness against model misspecification. For example, Politis (2003,2007a) explored model-free prediction in the practical setting of financial time series, i.e., a setting like example (2) with $\mu \equiv 0$ and a parametric structure for σ, and found that the model-free predictors *outperform* the ones based on the popular ARCH/GARCH models.

In this paper, we identify the underlying principles and elements of model-free prediction that apply equally to cases where the breakdown of the i.i.d. assumption is either due to non-identical distributions, i.e., the regression example (1), and/or due to dependence in the data as in example (2). In Section 2, these general principles for model-free prediction are theoretically formulated; their essence is based on the notion of transforming a given setup into one that is easier to work with, e.g., i.i.d. or Gaussian. We also describe how

the model-free prediction principle can be combined with the bootstrap to yield frequentist predictive distributions in a very general framework.

The remainder of the paper is devoted to the regression example (1) that is quintessential in statistical practice. Model-based and model-free predictors are derived in detail in Sections 3 and 4 respectively, with particular emphasis on the derivation of predictive distributions and intervals. As a running example we use the Canadian earnings data from the 1971 Canadian Census; this is a wage vs. age dataset concerning 205 male individuals with high-school education. Finite-sample simulations are also provided comparing the different prediction intervals in the context of nonparametric, as well as linear, regression. In the latter case, a model-free variation on the model-based theme seems to give a long awaited answer on the well-known problem of *under*coverage of bootstrap prediction intervals.

2 Model-free prediction: a basic principle

2.1 The i.i.d. case

As already mentioned, the prediction problem is most interesting in cases where the i.i.d. assumption breaks down. However, we now briefly focus on the i.i.d. case in order to motivate the more general case.

Consider real-valued data $Y_1, \ldots, Y_n$ i.i.d. from the (unknown) distribution F_Y. The goal is prediction of a future value Y_{n+1} based on the data. It is apparent that F_Y *is* the predictive distribution, and its quantiles could be used to form predictive intervals. Furthermore, different measures of center of location of the distribution F_Y can be used as (point) predictors of Y_{n+1}. In particular, the mean and median of F_Y are of interest since they represent optimal predictors under an L_2 and L_1 criterion respectively.

Of course, F_Y is unknown but can be estimated by the empirical distribution of the data $Y_1, \ldots, Y_n$ denoted by $\hat{F}_Y$. Thus, practical model-free predictive intervals will be based on quantiles of $\hat{F}_Y$, and the L_2 and L_1 optimal predictors will be approximated by the mean and median of $\hat{F}_Y$ respectively.

2.2 The general prediction paradigm

In general, the data $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ may not be i.i.d. so the predictive distribution of Y_{n+1} given the data may depend on $\underline{Y}_n$ and on $\mathbf{X}_{n+1}$ which is a matrix of observable, explanatory (predictor) variables; for concreteness, we will assume the predictors are deterministic but provisions for random regressors can be made. The notation $\mathbf{X}_n$ here is cumulative, i.e., $\mathbf{X}_n$ is the collection of all predictor variables associated with the data $\underline{Y}_t$ for $t = 1, \ldots, n$; in the regression example of eq. (1), the matrix $\mathbf{X}_n$ would be formed by

concatenating together all the (fixed-length) predictor vectors $\underline{x}_t, t = 1, \ldots, n$.

Let Y_t take values in the linear space $\mathbf{B}$ which typically will be $\mathbf{R}^d$ for some integer d. The goal is to predict $g(Y_{n+1})$ based on $\underline{Y}_n$ and $\mathbf{X}_{n+1}$ *without* invoking any particular model; here g is some real-valued (measurable) function on $\mathbf{B}$. The key to successful model-free prediction is the following *model-free prediction principle* that was first presented in a conference announcement (extended abstract) of Politis (2007b). Intuitively, the basic idea is to transform the non-i.i.d. setup to an i.i.d. dataset for which prediction is easy—even trivial—, and then transform back to the original setting to obtain the model-free prediction.

Model-free prediction principle.
(a) For any integer $m \geq$ some m_o, suppose that a transformation H_m is found that maps the data $\underline{Y}_m = (Y_1, \ldots, Y_m)'$ and the explanatory variables $\mathbf{X}_m$ onto the vector $\underline{\epsilon}_m^{(m)} = (\epsilon_1^{(m)}, \ldots, \epsilon_m^{(m)})'$ where the $\{\epsilon_i^{(m)}, i = 1, \ldots, m\}$ are i.i.d. *with distribution F_m, and F_m is such that $F_m \stackrel{\mathcal{L}}{\Longrightarrow}$ some F as $m \to \infty$.*
(b) Suppose that the transformation H_m is invertible for all m (possibly modulo some initial conditions denoted by IC), and—in particular—that one can solve for Y_m in terms of $\underline{Y}_{m-1}, \mathbf{X}_m$, and $\epsilon_m^{(m)}$ alone, i.e., that

$$Y_m = g_m(\underline{Y}_{m-1}, \mathbf{X}_m, \epsilon_m^{(m)}) \tag{3}$$

and

$$\underline{Y}_{m-1} = f_m(\underline{Y}_{m-2}, \mathbf{X}_m;\ \epsilon_1^{(m)}, \ldots, \epsilon_{m-1}^{(m)};\ IC) \tag{4}$$

for some functions g_m and f_m and for all $m \geq m_o$.
(c) Then, the L_2-optimal model-free predictor of $g(Y_{n+1})$ on the basis of the data $\underline{Y}_n$ and the predictors $\mathbf{X}_{n+1}$ is given by the (conditional) expectation
$\int G_{n+1}(\underline{Y}_n, \mathbf{X}_{n+1}, \epsilon) dF_{n+1}(\epsilon)$ where $G_{n+1} = g \circ g_{n+1}$ denotes composition of functions.
(d)The whole predictive distribution of $g(Y_{n+1})$ is given by the distribution of the random variable $G_{n+1}(\underline{Y}_n, \mathbf{X}_{n+1}, \epsilon_{n+1})$ where ϵ_{n+1} is drawn from distribution F_{n+1} and is independent to $\underline{Y}_n$. The median of this predictive distribution yields the L_1-optimal model-free predictor of $g(Y_{n+1})$ given $\underline{Y}_n$ and $\mathbf{X}_{n+1}$.

Parts (c) and (d) above outline a general approach to the problem of prediction of (a function of) Y_{n+1} given a dataset of size n. As will be apparent in the sequel, the application of Model-free prediction hinges on the aforementioned transformation H_m and its inverse for $m = n$ and $m = n + 1$.

The predictive distribution in part (d) above is meant to be *conditional* on the value of $\underline{Y}_n$ (and the value of $\mathbf{X}_{n+1}$ when the latter is random), as is the expectation in part (c). Note also the tacit understanding that the 'future' ϵ_{n+1} is independent to the conditioning

variable $\underline{Y}_n$; this assumption is directly implied by eq. (4) which itself—under some assumptions on the function g_m—could be obtained by iterating (back-solving) eq. (3). The presence of initial conditions such as IC in eq. (4) is familiar in time series problems of autoregressive nature where IC would typically represent values $Y_0, Y_{-1}, \ldots, Y_{-p}$ for a finite value p; the effect of the initial conditions is negligible for large n. In regression problems the presence of initial conditions would not be required if the regression errors can be assumed to be independent as in eq. (1).

Fact 2.1 *Under regularity conditions, a transformation such as H_m of part (a) 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 with F_m being Uniform (0,1). 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.*

To briefly explain the above, 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. To elaborate, for $k > 1$ define the conditional distributions $D_k(y_k|y_{k-1}, \ldots, y_1) = P\{Y_k \leq y_k|Y_{k-1} = y_{k-1}, \ldots, Y_1 = y_1\}$, and let $D_1(y_1) = P\{Y_1 \leq y_1\}$. Then the Rosenblatt transformation amounts to letting $U_1 = D_1(Y_1), U_2 = D_2(Y_2|Y_1), U_3 = D_3(Y_3|Y_2, Y_1), \ldots$, and $U_m = D_m(Y_m|Y_{m-1}, \ldots, Y_2, Y_1)$. The problem is that the conditional distributions D_k for $k \geq 1$ are typically unknown and must be estimated (in a continuous fashion) from the $\underline{Y}_m$ data at hand. It is apparent that unless there is some additional structure, e.g. Markovian or regression as in Section 4, this estimation task may be unreliable or even infeasible for large k. As an extreme example, note that to estimate D_m one would have only one point (in m-dimensional space) to work with; thus, without additional assumptions, the estimate of D_m would be a point mass which is (a) a completely unreliable estimate, and (b) it is of little use in terms of constructing a probability integral trasform due to its discontinuity.

Remark 2.1 Eq. (3) with $\epsilon_i^{(m)}$ being i.i.d. from distribution F_m looks like a model equation but it is more general than a typical model. For one thing, the functions g_m and F_m may change with m, and so does $\epsilon_i^{(m)}$ which, in essense, is a triangular array of i.i.d. random variables. Furthermore, no assumptions are made *a priori* on the form of g_m. However, the process of starting without a model, and—by this transformation technique—arriving at a model-like equation deserves the name *model-free model-fitting*, (MF^2 for short).

Remark 2.2 The predictive distribution in part (d) above is the *true* distribution in this setup but it is unusable as such since it depends on many potentially unknown quantities. For example, the distribution F_{n+1} will typically be unknown but it can be consistently estimated by $\hat{F}_n$, the empirical distribution of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$, which can then be plugged-in to compute estimates of the aforementioned (conditional) mean, median, and predictive distribution. Similarly, if the form of function g_{n+1} is unknown, a consistent estimator $\hat{g}_{n+1}$ could be plugged-in. The resulting empirical estimates of the (conditional) mean and median would typically be quite accurate but such a 'plug-in' empirical estimate of the predictive distribution will be too narrow, i.e., possessing a smaller variance and/or inter-quartile range than ideal. The correct predictive distribution should incorporate the variability of $\hat{F}_n$ and/or $\hat{g}_{n+1}$. The only general *frequentist* way to nonparametrically capture such a widening of the predictive distribution via *resampling*; see Section 2.6 for more details.

2.3 Transformation into Gaussianity as a stepping stone

The prediction principle sounds deceptively simple but its application is not. The task of finding a set of candidate transformations H_n for any given particular setup is challenging, and demands expertise and ingenuity; see Remark 2.3 and Section 2.5 for some discussion to that effect. Once, however, a set of candidate transformations is identified (and denoted by $\mathcal{H}$), the procedure is easy to delineate: *Choose the transformation $H_n \in \mathcal{H}$ that minimizes the (pseudo)distance $d(\mathcal{L}(H_n(\underline{Y}_n)), \mathcal{F}_{iid,n})$ over all $H_n \in \mathcal{H}$;* here $\mathcal{L}(H_n(\underline{Y}_n))$ is the probability law of $H_n(\underline{Y}_n)$, and $\mathcal{F}_{iid,n}$ is the space of all distributions associated with an n-dimensional random vector whose $\mathbf{B}$–valued coordinates are i.i.d., i.e., the space of all distributions of the type $F \times F \times \cdots \times F$ where F is an arbitrary distribution on space $\mathbf{B}$. There are many choices for the (pseudo)distance d; see Hong and White (2005) and the references therein.

Remark 2.3 If a model such as (1) or (2) is plausible, then the model itself suggests the form of the transformation H_n, and the residuals from model-fitting would serve as the 'transformed' values $\epsilon_t^{(n)}$. Of course, the goodness of the model should now be assessed in terms of achieved "i.i.d."—ness of these residuals. It is relatively straightforward—via the usual graphical methods—to check that the residuals have identical distributions but checking their independence is trickier; see e.g. Hong (1999). However, if the residuals happened to be (jointly) Gaussian, then checking their independence is easy since it is equivalent to checking for correlation, e.g. portmanteau test, Ljung-Box, etc. This observation motivates the following variation of the prediction principle that is particularly useful in the case of dependent data:

(a′) For any m, find a transformation H_m on $\mathbf{B}^m$ that maps the data $\underline{Y}_m = (Y_1, \ldots, Y_m)'$ into a Gaussian vector $\underline{W}_m^{(m)} = (W_1^{(m)}, \ldots, W_m^{(m)})'$ having covariance matrix V_m.
(b′) Use a linear transformation to map $\underline{W}_m^{(m)}$ into the i.i.d. Gaussian vector $\underline{\epsilon}_m^{(m)} = (\epsilon_1^{(m)}, \ldots, \epsilon_m^{(m)})'$, and then continue with parts (c) and (d) of the prediction principle.

In applications, the above linear trasformation may be estimated by fitting a linear model and/or by direct estimation of the covariance matrix V_n from the transformed data $W_1^{(n)}$, $\ldots, W_n^{(n)}$ using some extra assumption on its structure, e.g., a Toeplitz structure in stationary time series as in McMurry and Politis (2010), or an appropriate shrinkage/regularization technique as in Bickel and Li (2006); then, the estimate $\hat{V}_n$ must be extrapolated to give an estimate of V_{n+1}.

Normalization as a prediction 'stepping stone' can be formalized in much the same way as before. To elaborate, once $\mathcal{H}$, the set of candidate transformations is identified, the procedure is to: *choose the transformation $H_n \in \mathcal{H}$ that minimizes the distance $d(\mathcal{L}(H_n(\underline{Y}_n)), \Phi_n)$ over all $H_n \in \mathcal{H}$* where now Φ_n is the space of all n-dimensional Gaussian distributions on $\mathbf{B}$. Many choices for the distance d are again available, including usual goodness-of-fit favorites such as the Kolmogorov-Smirnov or χ^2 test; a pseudo-distance based on the Shapiro-Wilk statistic is also a valid alternative here. Interestingly, in the setting of financial data, i.e., a heteroskedastic time series setup like example (2) with $\mu \equiv 0$ and heavy tails, Politis (2003,2007a) was able to achieve normalization by a kurtosis-based distance measure.

Remark 2.4 Now that H_n is essentially a *normalizing* transformation, a collection of graphical and exploratory data analysis (EDA) tools are also available. Some of these tools include: (a) Q-Q plots of the $W_1^{(n)}, \ldots, W_n^{(n)}$ data to test for Gaussianity; (b) Q-Q plots of linear combinations of $W_1^{(n)}, \ldots, W_n^{(n)}$ to test for *joint* Gaussianity; and (c) autocorrelation plots of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$ to test for independence—since in the (jointly) Gaussian case, independence is tantamount to zero correlation.

2.4 Comparison with other approaches

The application of the prediction principle appears similar in spirit to the Minimum Distance Method (MDM) of Wolfowitz (1957). Nevertheless, their objectives are quite different since MDM is typically employed for parameter estimation and testing whereas in the prediction paradigm there is no interest in parameters. A typical MDM searches for the parameter $\hat{\theta}$ that minimizes the distance $d(\hat{F}_n, \mathcal{F}_\theta)$, i.e., the distance of the empirical distribution $\hat{F}_n$ to a parametric family $\mathcal{F}_\theta$. In this sense, it is apparent that MDM sets an ambitious target (the parametric family $\mathcal{F}_\theta$) but there is no necessity of actually 'hitting' this target. By contrast, the prediction principle sets the minimal target of independence but its successful application requires that this minimal target is more or less achieved.

In anticipation of the detailed discussion on the setup of regression in Sections 3 and 4, it should be mentioned that devising transformations in regression has always been thought to be a crucial issue that received attention early on by statistics pioneers such as F. Anscombe, M.S. Bartlett, R.A. Fisher, etc.; see the excellent exposition of DasGupta (2008, Ch. 4) and the references therein, as well as Draper and Smith (1998, Ch. 13), Atkinson (1985), and Carroll and Ruppert (1988).

Regarding nonparametric regression in particular, the power family of Box and Cox (1964) has been routinely used in practice, as well as more elaborate, computer-intensive transformation techniques. Of the latter, we single out the ACE algorithm of Breiman and Friedman (1985), and the AVAS transformation of Tibshirani (1988). Both ACE and AVAS are very useful for transforming the data in a way that the usual additive nonparametric regression model is applicable with AVAS also achieving variance stabilization. However, as will be apparent in Section 4, the model-free approach to nonparametric regression is *insensitive* to whether such pre-processing by Box/Cox, ACE or AVAS has taken place. Consequently, the model-free practitioner is relieved from the need to find an optimal transformation; thus, model-free model-fitting in regression is a totally *automatic* technique.

2.5 Model-free model-fitting in practice

As mentioned in Section 2.3, the task of identifying the transformation H_n for a given particular setup is expected to be challenging since it is analogous to the difficult task of identifying a good model for the data at hand, i.e., model-building. Thus, faced with a new dataset, the model-free practitioner could/should take advantage of all the model-building know-how associated with the particular problem. The resulting 'best' model can then serve as the starting point in concocting the desired transformation as mentioned in Remark 2.3.

As with model-building, the candidate transformation will typically depend on some unknown parameter, say θ, that may be finite-dimensional or infinite-dimensional—the latter corresponding to a 'nonparametric' model. There are many potential strategies for chosing an optimal value for the parameter θ based on the data; the simplest strategy is to:

(A) Continue with the model-building analogy, and use standard estimation techniques such as Maximum Likelihood (ML) or Least Squares (LS) when θ is finite-dimensional, or standard nonparametric/smoothing techniques when θ is infinite-dimensional.

If step (A) is not successful in rendering the transformed data i.i.d., then the strategy may be modified as follows.

(B) The parameter θ may be divided in two parts, i.e., $\theta=(\theta_1, \theta_2)$ where θ_1 is of finite(and hopefully small) dimension. Firstly, (θ_1, θ_2) are fitted using standard methods as

in strategy (A). Then, using the fitted value for θ_2, a new search for θ_1 is initiated choosing the θ_1 value that renders the transformed data closest to being i.i.d.

Nevertheless, in certain examples the form of the desired transformation H_n is apparent; this is—fortunately—the case in the regression example whether an additive model is true (Section 3) or not (Section 4).

Remark 2.5 It has been noted that the model-free (MF) approach relinquishes the notion of a model only to replace it with that of a transformation; indeed, the MF approach could equally be termed a *Transformation-based approach to inference*. To further elucidate the similarities and differences between the MF and the model-based (MB) approaches, consider a setup where an additive model with respect to i.i.d. errors is indeed available, e.g., assume $Y_t = \mu(x_t) + \varepsilon_t$ with ε_t being i.i.d. $(0, \sigma^2)$, and $\mu(\cdot)$ an unknown function; this is the setup that will be analyzed more generally in Section 3. It is apparent that in order to concoct a transformation towards i.i.d.–ness, the MF practitioner would do well by estimating the mean $\mu(x_t)$ and subtracting it from the Y_t data. Hence, when a model with respect to i.i.d. errors is available, the MF practitioner may indeed proceed in a similar way as in the model-based approach.[1] Interestingly, the Model-Free principle appears to be more primitive than Least Squares, i.e., implying Least Squares (or even L_1 regression) under certain conditions—see e.g. Section 3.6 of Politis (2010); so using (say) a LS estimator of $\mu(\cdot)$ is very much in line with the MF principle. Of course, when a model is *not* available, the MF approach has little competition—see Section 4 in what follows.

2.6 Model-free predictive distributions and resampling

As mentioned in Remark 2.2, plugging-in estimates of $\hat{F}_n$ and/or $\hat{g}_{n+1}$ in the theoretical predictive distribution of the model-free principle may result in an estimated predictive distribution that is too narrow. The only general frequentist way to practically correct for that is via *resampling*. Fortunately, the model-free principle is ideally amenable to the i.i.d. bootstrap of Efron (1979).

For simplicity—and concreteness—we assume henceforth that the effect of the initial conditions IC is negligible as is, e.g., in the regression example (1). We will focus on constructing bootstrap prediction integrals of the *'root'* type in analogy to the well-known confidence interval construction; cf. Hall (1992), Efron and Tibshirani (1993), Davison and Hinkley (1997), or Shao and Tu (1995). To see how, let $\Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$

[1]The qualitative difference is that the interest of the MF practitioner is on observable quantities, i.e., current and future data, as opposed to unobservable model parameters and estimates thereof. In this sense, despite being frequentist in nature, the MF principle is in concordance with Bruno de Finetti's statistical philosophy—see e.g. Dawid (2004) and the references therein.

denote the best (with respect to either L_1 or L_2) data-based *point predictor* of $g(Y_{n+1})$ as obtained by the Model-free prediction principle coupled with Remark 2.2. The notation $\Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$ is meant to clarify how the point predictor depends on known (given) vs. estimated quantities; for example, $\hat{F}_n$ is the empirical distribution of $\varepsilon_1^{(n)}, ..., \varepsilon_n^{(n)}$, and $\hat{g}_{n+1}$ is the estimated prediction function associated with the estimated transformation $\hat{H}_n$. To elaborate, the L_2–optimal point predictor of $g(Y_{n+1})$ is given by:

$$\Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n) = \int g\left(\hat{g}_{n+1}(\underline{Y}_n, \mathbf{X}_{n+1}, \varepsilon)\right) d\hat{F}_n(\varepsilon) = \frac{1}{n}\sum_{j=1}^{n} g\left(\hat{g}_{n+1}(\underline{Y}_n, \mathbf{X}_{n+1}, \varepsilon_j^{(n)})\right);$$

similarly, the L_1–optimal predictor is the median of the set $\{g\left(\hat{g}_{n+1}(\underline{Y}_n, \mathbf{X}_{n+1}, \varepsilon_j^{(n)})\right)$, for $j = 1, ..., n\}$. Then, our 'root' is nothing else than the prediction error:

$$g(Y_{n+1}) - \Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n) \tag{5}$$

whose distribution can be approximated by that of the bootstrap root:

$$g(Y^*_{n+1}) - \Pi(g, \hat{g}^*_{n+1}, \underline{Y}^*_n, \mathbf{X}_{n+1}, \hat{F}^*_n) \tag{6}$$

where $\hat{g}^*_{n+1}, \hat{F}^*_n$ and $\underline{Y}^*_n$ are bootstrap quantities to be formally defined in step 2 of the Resampling Algorithm that is outlined below.

Resampling algorithm for model-free predictive distribution of $g(Y_{n+1})$

1. Based on the data $\underline{Y}_n$, estimate the transformation H_n and its inverse H_n^{-1} by $\hat{H}_n$ and $\hat{H}_n^{-1}$ respectively. In addition, estimate g_{n+1} by $\hat{g}_{n+1}$.
2. Use $\hat{H}_n$ to obtain the transformed data, i.e., $(\varepsilon_1^{(n)}, ..., \varepsilon_n^{(n)}) = \hat{H}_n(\underline{Y}_n)$. By construction, the data $\varepsilon_1^{(n)}, ..., \varepsilon_n^{(n)}$ are approximately i.i.d.
   (a) Sample randomly (with replacement) the data $\varepsilon_1^{(n)}, ..., \varepsilon_n^{(n)}$ to create the bootstrap pseudo-data $\varepsilon_1^\star, ..., \varepsilon_n^\star$ whose empirical distribution is denoted $\hat{F}_n^\star$.
   (b) Use the inverse transformation $\hat{H}_n^{-1}$ to create pseudo-data in the Y domain, i.e., let $\underline{Y}_n^\star = (Y_1^\star, ..., Y_n^\star) = \hat{H}_n^{-1}(\varepsilon_1^\star, ..., \varepsilon_n^\star)$.
   (c) Calculate a bootstrap pseudo-response Y^*_{n+1} as the point $\hat{g}_{n+1}(\underline{Y}_n^\star, \mathbf{X}_{n+1}, \varepsilon)$ where ε is drawn randomly from the set $(\varepsilon_1^{(n)}, ..., \varepsilon_n^{(n)})$.
   (d) Based on the pseudo-data $\underline{Y}_n^\star$, estimate the function g_{n+1} by $\hat{g}_{n+1}^\star$ respectively.
   (e) Calculate a bootstrap root replicate using eq. (6).

3. Steps (a)—(e) in the above should be repeated a large number of times (say B times), and the B bootstrap root replicates should be collected in the form of an empirical distribution whose α—quantile is denoted by $q(\alpha)$.

4. A $(1-\alpha)100\%$ *equal-tailed* predictive interval (of root type) for $g(Y_{n+1})$ is given by

$$[\Pi + q(\alpha/2),\ \Pi + q(1-\alpha/2)] \tag{7}$$

where Π is short-hand for $\Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$.

5. Finally, our model-free estimate of the predictive distribution of $g(Y_{n+1})$ is the empirical distribution of bootstrap roots obtained in step 3 *shifted to the right* by the number Π; this is equivalent to the empirical distribution of the B bootstrap root replicates when the quantity Π is added to each. [Recall that the predictive distribution of $g(Y_{n+1})$ is—by definition—conditional on $\underline{Y}_n$ and $\mathbf{X}_{n+1}$; hence, the quantity $\Pi = \Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$ is a constant given $\underline{Y}_n$ and $\mathbf{X}_{n+1}$.]

The above algorithm is closely related to so-called 'residual bootstrap' schemes in model-based situations—cf. Efron (1979). The only difference is that, in the model-free setting, the i.i.d. variables $\varepsilon_1^{(n)}, ..., \varepsilon_n^{(n)}$ are not residuals but the outcome of the data-transformation.

Note that, using an estimate of the prediction error variance, prediction intervals of the *studentized* root type can also be constructed—see Section 2.6 of Politis (2010). However, in contrast to what happens in confidence intervals, studentization does not ensure second order accuracy of prediction intervals; see e.g. Shao and Tu (1995, Ch. 7.3).

3 Model-based prediction in regression

3.1 Model-based nonparametric regression

We now focus on the nonparametric regression model of eq. (1). For simplicity, the regressor $\underline{x}_t$ is assumed univariate and deterministic, and denoted simply as x_t; the case of a multivariate regressor is handled in an identical fashion although, of course, the *caveat* of the curse of dimensionality must always be born in mind. Thus, throughout Section 3, our data $\{(Y_t, x_t), t = 1, ..., n\}$ are assumed to have been generated by the model

$$Y_t = \mu(x_t) + \sigma(x_t)\ \varepsilon_t\ ,\quad t = 1, \ldots, n, \tag{8}$$

with ε_t being i.i.d. (0,1) from the (unknown) distribution F; the functions $\mu(\cdot)$ and $\sigma(\cdot)$ are also unknown but assumed to possess some degree of smoothness (differentiability, etc.).

There are many approaches towards nonparametric estimation of the functions μ and σ, e.g., wavelets and orthogonal series, smoothing splines, local polynomials, and kernel

smoothers. The reviews by Altman (1992) and Schucany (2004) give concise introductions to popular methods of nonparametric regression with emphasis on kernel smoothers; book-length treatments are given by Härdle (1990), Hart (1997), Fan and Gijbels (1996), and Loader (1999). The above references focus on estimation of the conditional mean (and other moments). Regarding estimation of conditional quantiles, the book by Koenker (2005) is an excellent reference, and includes a chapter on boostrapping quantile regression estimators; see alo Gangopadhyay and Sen (1990), Hahn (1995), and Horowitz (1998) to that effect.

For simplicity of presentation, we will focus here on nonparametric regression based on kernel smoothing. Nevertheless, it is important to stress that the prediction inference procedures of this paper can equally be implemented with *any* other appropriate regression estimator, be it of parametric or nonparametric form. A popular—and very intuitive—form of a kernel smoother is the Nadaraya-Watson estimator (Nadaraya (1964), Watson (1964)) defined by

$$m_x = \sum_{i=1}^{n} Y_i \tilde{K}\left(\frac{x - x_i}{h}\right) \tag{9}$$

where h is the bandwidth, $K(x)$ is a symmetric kernel function with $\int K(x)dx = 1$, and

$$\tilde{K}\left(\frac{x - x_i}{h}\right) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x-x_k}{h}\right)}. \tag{10}$$

Similarly, the Nadaraya-Watson estimator of $\sigma(x)$ is given by s_x where

$$s_x^2 = M_x - m_x^2 \quad \text{where} \quad M_x = \sum_{i=1}^{n} Y_i^2 \tilde{K}\left(\frac{x - x_i}{q}\right), \tag{11}$$

and q is another bandwidth parameter. Selection of the bandwidth parameters h and q is often done by *cross-validation*. To elaborate, let e_t denote the *fitted* residuals, i.e.,

$$e_t = (Y_t - m_{x_t})/s_{x_t} \quad \text{for } t = 1, \ldots, n. \tag{12}$$

and $\tilde{e}_t$ the *predictive* residuals, i.e.,

$$\tilde{e}_t = \frac{Y_t - m_{x_t}^{(t)}}{s_{x_t}^{(t)}}, \quad t = 1, \ldots, n \tag{13}$$

where $m_x^{(t)}$ and $M_x^{(t)}$ denote the estimators m and M respectively computed from the *delete-Y_t* dataset: $\{(Y_i, x_i),\ i = 1, \ldots, t-1 \text{ and } i = t+1, \ldots, n\}$, and evaluated at the point x; as usual, we define $s_{x_t}^{(t)} = \sqrt{M_{x_t}^{(t)} - (m_{x_t}^{(t)})^2}$. In other words, $\tilde{e}_t$ is the (standardized) error in trying to predict Y_t from the aforementioned delete-Y_t dataset.

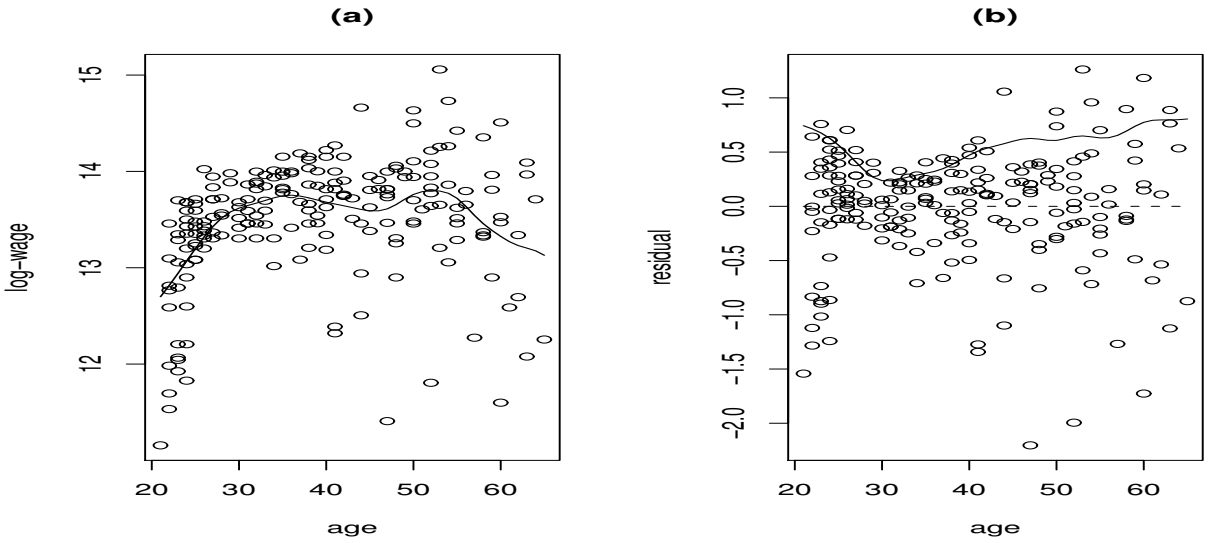

Figure 1: (a) Log-wage vs. age data with fitted kernel smoother m_x (solid line). (b) Plot of the unstudentized residuals $Y - m_x$ with superimposed estimated standard deviation s_x (solid line).

Cross-validation amounts to picking the bandwidths[2] that minimize PRESS $= \sum_{t=1}^{n} \tilde{e}_t^2$, i.e., the PREdictive Sum of Squared residuals. PRESS is an L_2 measure that is obviously non-robust in case of heavy-tailed errors and/or outliers. For this reason, in this paper, we favor cross-validation based on an L_1 criterion. It is more robust, and is not any more computationally expensive than PRESS—see Appendix B of Politis (2010) for more details. L_1—cross-validation amounts to picking the bandwidths that minimize $\sum_{t=1}^{n} |\tilde{e}_t|$; the latter could be denoted PRESAR, i.e., PREdictive Sum of Absolute Residuals. Note that L_1—cross-validation imposes the L_1 penalty on the predictive residuals, and thus should be distinguished from Tibshirani's (1996) Lasso that imposes an L_1 penalty on the regression parameters.

[2]Rather than doing a two-dimensional search over h and q to minimize PRESS, the simple constraint $q = h$ will be imposed in what follows which has the additional advantage of rendering $M_x \geq m_x^2$ as needed for a well-defined estimator s_x^2 in eq. (11). Note, that the choice $q = h$ is not necessarily optimal; see e.g. Wang et al. (2008). Furthermore, note that these are global bandwidths; techniques for picking *local* bandwidths, i.e., a different optimal bandwidth for each x, are widely available but will not be discussed further here in order not to obscure the paper's main focus. Similarly, there are several recent variations on the cross-validation theme such as the one-sided cross-validation of Hart and Yi (1998), and the far casting cross-validation for dependent data of Carmack et al. (2009) that present attractive alternatives. However, our discussion will focus on the well-known standard form of cross-validation for concreteness especially since our aim is to show how the Model-Free prediction principle applies in nonparametric regression with any type of kernel smoother, and any type of bandwidth selector.

Remark 3.1 If there are large 'gaps' in the scatterplot of the data, i.e., if there are large x–regions within the range of $x_1, ..., x_n$ where no data are available, then a 'local' bandwidth, i.e., a bandwidth that depends on x, or a *k–nearest neighbor* technique may be used; see e.g. Li and Racine (2007, Ch. 14).

As a running example we use the Canadian high-school graduate earnings data from the 1971 Canadian Census; this is a wage vs. age dataset concerning 205 male individuals with common education (13th grade). The data are available under the name `cps71` within the `np` package of R, and are discussed in Pagan and Ullah (1999). Figure 1 (a) presents a scatterplot of the data with the fitted kernel estimator m_x superimposed using a normal kernel. The kernel smoother seems to be problematic at the left boundary. The problem can be alleviated either using a local linear smoother as in Figure 2 of Schucany (2004), or by employing the reflection technique of Hall and Wehrly (1991); see also the recent paper by Dai and Sperlich (2010) for a comparison of different boundary correction techniques for kernel smoothers. We will not elaborate further here on these issues since our focus is on the general Model-free Prediction method which can equally be implemented with *any* chosen regression estimator. Finally, Figure 1 (b) shows a scatterplot of the unstudentized residuals $Y - m_x$ with the estimated standard deviation s_x superimposed.

3.2 Model-based prediction in regression

The prediction problem amounts to predicting the future response Y_{f} associated with a potential design point x_{f}. Recall that the L_2–optimal (point) predictor of Y_{f} is $E(Y_{\mathrm{f}}|x_{\mathrm{f}})$, i.e., the expected value [3] of the response Y_{f} associated with design point x_{f}. Under model (8), we have that $E(Y_{\mathrm{f}}|x_{\mathrm{f}}) = \mu(x_{\mathrm{f}})$. However, if the Y_t–data are heavy-tailed, the L_1–optimal predictor might be preferred; this would be given by the *median* response Y_{f} associated with design point x_{f}; under model (8), this is given by $\mu(x_{\mathrm{f}}) + \sigma(x_{\mathrm{f}}) \cdot median(F)$. If the error distribution F is symmetric, then the L_2— and L_1—optimal predictors coincide.

To obtain practically useful predictors, the unknown quantities $\mu(x), \sigma(x)$ and $median(F)$ must be estimated and plugged in the formulas of optimal predictors. Naturally, $\mu(x_{\mathrm{f}})$ and $\sigma(x_{\mathrm{f}})$ are estimated by $m_{x_{\mathrm{f}}}$ and $s_{x_{\mathrm{f}}}$ of eq. (9) and (11). The unknown F can be estimated by $\hat{F}_e$, the empirical distribution of the residuals $e_1, \ldots, e_n$ that are defined in eq. (12). Hence, the practical L_2— and L_1—optimal *model-based* predictors of Y_{f} are given respectively by $\widehat{Y}_{\mathrm{f}} = m_{x_{\mathrm{f}}}$ and $\widetilde{Y}_{\mathrm{f}} = m_{x_{\mathrm{f}}} + s_{x_{\mathrm{f}}} \cdot median(\hat{F}_e)$.

Suppose, however, that our objective is predicting the future value $g(Y_{\mathrm{f}})$ associated with design point x_{f} where $g(\cdot)$ is a function of interest; this possibility is of particular importance

[3] In general, the L_2–optimal predictor of Y_{f} would be given by the conditional expectation of Y_{f} given $Y_1, \ldots, Y_n$ as well as x_{f}; see e.g. Goldberger (1962). However, under model (8), the Y data are assumed independent; therefore, $E(Y_{\mathrm{f}}|x_{\mathrm{f}}, Y_1, \ldots, Y_n)$ simplifies to just $E(Y_{\mathrm{f}}|x_{\mathrm{f}})$.

due to the fact that data transformations such as Box/Cox, ACE, AVAS, etc. are often applied in order to arrive at a reasonable additive model such as (8). For example, the wages in dataset `cps71` have been logarithmically transformed before model (8) was fitted in Figure 1 (a); in this case, $g(x) = \exp(x)$ since naturally we are interested in predicting wage not log-wage.

In such a case, the model-based L_2–optimal (point) predictor of $g(Y_{\rm f})$ is $E(g(Y_{\rm f})|x_{\rm f})$ which can be estimated by $n^{-1}\sum_{i=1}^{n} g\left(m_{x_{\rm f}} + \sigma_{x_{\rm f}} e_i\right)$. Unfortunately, practitioners sometimes use the *naive* plug-in predictor $g(m_{x_{\rm f}})$ that can be grossly suboptimal since g is typically nonlinear. For instance, if g is convex, as in the exponential example above, Jensen's inequality immediately implies that the naive predictor $g(m_{x_{\rm f}})$ under-estimates its target, i.e., it is biased downward.

Similarly, the model-based L_1–optimal (point) predictor of $g(Y_{\rm f})$ can be approximated by the sample median of the set $\{g\left(m_{x_{\rm f}} + \sigma_{x_{\rm f}} e_i\right),\ i = 1, ..., n\}$; interestingly, the latter would be equivalent to the naive plug-in $g(\widetilde{Y}_{\rm f})$ as long as g is monotone.

3.3 A first application of the model-free prediction principle

Consider a dataset like the one depicted in Figure 1. Faced with this type of data, a practitioner may well decide to entertain a model like eq. (8) for his/her statistical analysis. However, even while fitting—and working with—model (8), it is highly unlikely that the practitioner will believe that this model is *exactly* true; more often than not, the model will be simply regarded as a convenient approximation.

Thus, in applying strategy (A) of Section 2.5, the model-free practitioner computes the fitted residuals $e_t = (Y_t - m_{x_t})/s_{x_t}$ that can be interpreted as an effort to center and studentize the $Y_1, \ldots, Y_n$ data. In this sense, they can be viewed as a preliminary transformation of the Y–data towards "i.i.d.–ness" since the residuals $e_1, \ldots, e_n$ have (approximately) same 1st and 2nd moment while the Y–data do not; see also Remark 2.5.

Recall that throughout Section 3 we have assumed that—possibly unbeknownst to the statistician—model (8) is true. Hence, the model-free practitioner should find (via the usual diagnostics) that to a good approximation the residuals $e_t = (Y_t - m_{x_t})/s_{x_t}$ from a model-based fit are close to being i.i.d.[4] However, the model-free practitioner does not see this as model confirmation, and may well try additional choices for centering and/or studentizing the data. Motivated by the studentizing transformation in Politis (2003,2007a), we consider a more general centering/studentization that may provide a better transformation for the

[4]Here, and for the remainder of Section 3, we will assume that the form of the estimator m_x is *linear* in the Y data; our running example of a kernel smoother obviously satisfies this requirement, and so do other popular methods such as local polynomial fitting.

model-free principle. Such a transformation is given by:

$$W_t = \frac{Y_t - \tilde{m}_{x_t}}{\tilde{s}_{x_t}} \ , \ t = 1, \ldots, n. \tag{14}$$

where

$$\tilde{m}_{x_t} = cY_t + (1-c)m_{x_t}^{(t)}, \ \tilde{M}_{x_t} = cY_t^2 + (1-c)M_{x_t}^{(t)} \ \text{ and } \ \tilde{s}_{x_t}^2 = \tilde{M}_{x_t} - \tilde{m}_{x_t}^2 \ . \tag{15}$$

In the above, $m_x^{(t)}$ and $M_x^{(t)}$ denote the estimators m and M respectively computed from the delete-Y_t dataset: $\{(x_i, Y_i),\ i = 1, \ldots, t-1 \text{ and } i = t+1, \ldots, n\}$, and evaluated at the point x. Note that the W's, as well as $\tilde{m}_{x_t}, \tilde{M}_{x_t}$, depend on the parameter $c \in [0,1)$ but this dependence is not explicitly denoted. The optimal choice of c will be discussed later. The case $c = 1$ is excluded as it leads to the trivial setting of $W_t = 0$, and an inconsistent $\tilde{m}_{x_t}$ that simply interpolates the data on the scatterplot; similarly problematic would be choosing c close to unity.

Nevertheless, eq. (14) is a general—and thus more flexible—reduction to residuals since it includes the fitted and predictive residuals as special cases. To see this, note that (9) implies that the choice $c = K(0)/\sum_{k=1}^{n} K\left(\frac{x_t - x_k}{h}\right)$ corresponds to $\tilde{m}_{x_t} = m_{x_t}$ and $\tilde{M}_{x_t} = M_{x_t}$ in which case eq. (14) reduces to eq. (12), i.e., the fitted residuals. Furthermore, consider the extreme case of $c = 0$; in this case, W_t is tantamount to a predictive residual, i.e., $W_t = \tilde{e}_t$ as defined in eq. (13).

Thus, eq. (14) is a good candidate for our search for a general transformation H_n towards "i.i.d.—ness" as the model-free prediction principle of Section 2 requires. With a proper choice of bandwidth (and the constant c), $W_1, \ldots, W_n$ would be—by construction—centered and studentized; hence, the first two moments of the W_t's are (approximately) constant. Since the original data are assumed independent, the W_t's are also approximately[5] independent. The (approximate) independence and constancy of the first two moments generally falls short of claiming that the W_t's are i.i.d. but it often suffices in practical work. Note, however, that the W_t's will be (approximately) i.i.d. here due to model (8) which is assumed to hold true.

3.4 Model-free/model-based prediction

Recall that the prediction problem amounts to predicting the future value Y_{f} associated with a potential design point x_{f}. As customary in a prediction problem one starts by investigating the distributional characteristics of the unobserved Y_{f} centered and studentized. To this

[5]Strictly speaking, the W_t's are not exactly independent because of dependence of m_{x_t} and s_{x_t} to m_{x_k} and s_{x_k}. However, under typical conditions, $m_x \xrightarrow{P} E(Y|x)$ and $s_x^2 \xrightarrow{P} Var(Y|x)$ as $n \to \infty$. Therefore, the W_t's are—at least—asymptotically independent.

effect, note that eq. (14) can still be written for the unobserved $Y_{\rm f}$, i.e., the yet unobserved $Y_{\rm f}$ is related to the yet unobserved $W_{\rm f}$ by

$$W_{\rm f} = \frac{Y_{\rm f} - \tilde{m}^{\rm f}_{x_{\rm f}}}{\tilde{s}^{\rm f}_{x_{\rm f}}} \tag{16}$$

where $\tilde{m}^{\rm f}$ and $\tilde{s}^{\rm f}$ are the estimators from eq. (9) and (11) but computed from the *augmented* dataset that includes the full original dataset $\{(x_i, Y_i),\ i = 1, \ldots, n\}$ *plus* the pair $(x_{\rm f}, Y_{\rm f})$. As in eq. (15) we have:

$$\tilde{m}^{\rm f}_{x_{\rm f}} = cY_{\rm f} + (1-c)m_{x_{\rm f}},\ \tilde{M}^{\rm f}_{x_{\rm f}} = cY^2_{\rm f} + (1-c)M_{x_{\rm f}} \ \text{ and } \ \tilde{s}^{\rm f}_{x_{\rm f}} = \sqrt{\tilde{M}^{\rm f}_{x_{\rm f}} - (\tilde{m}^{\rm f}_{x_t})^2} \tag{17}$$

where $m_{x_{\rm f}}, M_{x_{\rm f}}$ are the estimators m, M computed from the original dataset as in Section 3.2 and evaluated at the candidate point $x_{\rm f}$.

Solving eq. (16) for $Y_{\rm f}$ is the key to model-free prediction as it would yield an equation like (3). As it turns out, the solution of eq. (16) is given by

$$Y_{\rm f} = m_{x_{\rm f}} + s_{x_{\rm f}} \frac{W_{\rm f}}{\sqrt{1 - c - cW^2_{\rm f}}}; \tag{18}$$

see Appendix A of Politis (2010) for details. Eq. (18) is the regression analog of the general eq. (3) of Section 2.2, and will form the basis for our model-free prediction procedure.

One may now ponder on the optimal choice of c. It is possible to opt to choose c with the goal of normalization of the empirical distribution of the W's in the spirit of the 'Gaussian stepping stone' of Section 2.3. But inasmuch as prediction is concerned, Gaussianity is not required. Since the W_t are (at least approximately) i.i.d., the model-free prediction principle can be invoked, and is equally valid for *any* value of c. It is interesting then to ask how the predictors based on eq. (18) depend on the value of c. Surprisingly (and thankfully), the answer is *not at all* ! To see this, note that after some algebra:

$$\frac{W_t}{\sqrt{1 - c - cW^2_t}} \equiv \tilde{e}_t \ \text{ for any } \ c \in [0, 1), \ \text{ and for all } \ t = 1, \ldots, n, \tag{19}$$

where the $\tilde{e}_t$s are the *predictive* residuals defined in eq. (13). In other words, the prediction equation (18) does *not* depend on the value of c, and can be simplified to:

$$Y_{\rm f} = m_{x_{\rm f}} + s_{x_{\rm f}} \tilde{e}_{\rm f}. \tag{20}$$

Eq. (20) will form the basis for our application of the model-free prediction principle under model (8). Since the model-free philosophy is implemented in a setup where model (8) is true, we will denote the resulting predictors by `MF/MB` to indicate both the model-free (MF) *construction*, as well as the predictor's model-based (MB) *realm of validity*.

To elaborate on the construction of MF/MB predictors, let $\hat{F}_{\tilde{e}}$ denote the empirical distribution of the predictive residuals $\tilde{e}_1, \ldots, \tilde{e}_n$. Then, the L_2— and L_1—optimal *model-free* predictors of the function $g(Y_\mathrm{f})$ are given, respectively, by the expected value and median of the random variable $g(Y_\mathrm{f})$ where Y_f as given in eq. (20) and $\tilde{e}_\mathrm{f}$ is a random variable drawn from distribution $\hat{F}_{\tilde{e}}$.

Focusing on the case $g(x) = x$, if follows that the L_2— and L_1—optimal MF/MB predictors of Y_f are given, respectively, by the expected value and median of the random variable given in eq. (20). Note, however, that the only difference between eq. (20) and the fitted regression equation $Y_t = m_{x_t} + s_{x_t} e_t$ as applied to the case where x_t is the future point x_f is the use of the predictive residuals $\tilde{e}_t$ instead of the regression residuals e_t. The different predictors are summarized in Table 3.1.

	Model-based	MF/MB case
Predictive equation	$Y_\mathrm{f} = m_{x_\mathrm{f}} + s_{x_\mathrm{f}} e_\mathrm{f}$	$Y_\mathrm{f} = m_{x_\mathrm{f}} + s_{x_\mathrm{f}} \tilde{e}_\mathrm{f}$
L_2—predictor of Y_f	m_{x_f}	$m_{x_\mathrm{f}} + s_{x_\mathrm{f}} \cdot \mathrm{mean}(\tilde{e}_i)$
L_1—predictor of Y_f	$m_{x_\mathrm{f}} + s_{x_\mathrm{f}} \cdot \mathrm{median}(e_i)$	$m_{x_\mathrm{f}} + s_{x_\mathrm{f}} \cdot \mathrm{median}(\tilde{e}_i)$
L_2—predictor of $g(Y_\mathrm{f})$	$n^{-1} \sum_{i=1}^{n} g\left(m_{x_\mathrm{f}} + \sigma_{x_\mathrm{f}} e_i\right)$	$n^{-1} \sum_{i=1}^{n} g\left(m_{x_\mathrm{f}} + \sigma_{x_\mathrm{f}} \tilde{e}_i\right)$
L_1—predictor of $g(Y_\mathrm{f})$	$\mathrm{median}(g\left(m_{x_\mathrm{f}} + \sigma_{x_\mathrm{f}} e_i\right))$	$\mathrm{median}(g\left(m_{x_\mathrm{f}} + \sigma_{x_\mathrm{f}} \tilde{e}_i\right))$

Table 3.1. Comparison of the model-based and MF/MB point prediction procedures obtained when model (8) is true.

3.5 Model-free/model-based prediction intervals

The model-based L_2—optimal predictor of Y_f from Table 3.1 uses the model information that the mean of the errors is exactly zero and does not attempt to estimate it. Another way of enforcing this model information is to center the residuals e_i to their mean, and use the centered residuals for prediction; centering the residuals was first pointed out by Freedman (1981) in a linear model setting.

The use of predictive residuals is both natural and intuitive since the objective is prediction. Furthermore, in case $\sigma^2(x)$ can be assumed to be constant,[6] simple algebra shows

$$\tilde{e}_t = e_t/(1 - \delta_{x_t}) \quad \text{where } \delta_{x_t} = K(0)/\sum_{k=1}^{n} K\left(\frac{x_t - x_k}{h}\right). \tag{21}$$

Eq. (21) suggests that the main difference between the fitted and predictive residuals is

[6] If $\sigma^2(x)$ is not assumed constant, then $\tilde{e}_t = e_t C_t/(1 - \delta_{x_t})$ where $C_t = s_{x_t}/s_{x_t}^{(t)}$.

their scale; their center should be about the same (and close to zero) since typically

$$\text{mean}(\tilde{e}_i) \approx 0 \;\text{ and median}(\tilde{e}_i) \approx 0. \tag{22}$$

Therefore, the model-based and MF/MB *point* predictors of Y_f are almost indistinguishable; this is, of course, reassuring since, when model (8) is true, the model-based procedures are obviously optimal. Nevertheless, due to the different scales of the fitted and predictive residuals, the difference between the two approaches is more pronounced in terms of construction of a predictive *distribution* for Y_f in which case the correct scaling of residuals is of paramount importance; see also the discussion in Section 3.6.

With regards to the construction of an accurate predictive distribution of Y_f, both approaches (model-based and MF/MB) are formally identical, the only difference being in the use of fitted vs. predictive residuals. The Resampling Algorithm of Section 2.6 reads as follows for the case at hand where the predictive function g_{n+1} is essentially determined by $\mu(x)$ and $\sigma(x)$.

RESAMPLING ALGORITHM FOR THE PREDICTIVE DISTRIBUTION OF $g(Y_\text{f})$

1. Based on the data $\underline{Y}_n$, construct the estimates m_x and s_x from which the fitted residuals e_i, and predictive residuals $\tilde{e}_i$ are computed for $i = 1, ..., n$.

2. For the model-based approach, let $r_i = e_i - n^{-1}\sum_j e_j$, for $i = 1, ...n$, whereas for the MF/MB approach, let $r_i = \tilde{e}_i$, for $i = 1, ...n$. Also let Π be a short-hand for $\Pi(g, m_x, s_x, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$, the chosen predictor from Table 3.1; e.g. for the L_2–optimal predictor we have $\Pi = n^{-1}\sum_{i=1}^n g\left(m_{x_\text{f}} + \sigma_{x_\text{f}} r_i\right)$

   (a) Sample randomly (with replacement) the data $r_1, ..., r_n$ to create the bootstrap pseudo-data $r_1^\star, ..., r_n^\star$ whose empirical distribution is denoted by $\hat{F}_n^\star$.

   (b) Create pseudo-data in the Y domain by letting $Y_i^\star = m_{x_i} + s_{x_i} r_i^\star$, for $i = 1, ...n$.

   (c) Calculate a bootstrap pseudo-response as $Y_\text{f}^\star = m_{x_\text{f}} + s_{x_\text{f}} r$ where r is drawn randomly from the set $(r_1, ..., r_n)$.

   (d) Based on the pseudo-data $Y_1^\star, ..., Y_n^\star$, re-estimate the functions $\mu(x)$ and $\sigma(x)$ by the kernel estimators $m_x^\star$ and $s_x^\star$ (with same kernel and bandwidths as the original estimators m_x and s_x).

   (e) Calculate a bootstrap root replicate as $g(Y_\text{f}^\star) - \Pi(g, m_x^*, s_x^*, \underline{Y}_n^*, \mathbf{X}_{n+1}, \hat{F}_n^*)$.

3. Steps (a)—(e) in the above are repeated B times, and the B bootstrap root replicates are collected in the form of an empirical distribution with α–quantile denoted by $q(\alpha)$.

4. Then, a $(1-\alpha)100\%$ equal-tailed predictive interval for $g(Y_{\rm f})$ is given by:

$$[\Pi + q(\alpha/2), \Pi + q(1-\alpha/2)]. \tag{23}$$

5. Finally, our estimate of the predictive distribution of $g(Y_{\rm f})$ is the empirical distribution of bootstrap roots obtained in step 3 shifted to the right by the number Π.

Fact 3.1 *When $\sigma^2(x)$ is constant, eq. (21) implies that $\delta_{x_t} > 0$, and thus $\tilde{e}_t$ will always be* larger *in absolute value (i.e., inflated) as compared to e_t. As a consequence, MF/MB prediction intervals will tend to be wider than their MB counterparts. Nevertheless, this difference disappears asymptotically since $\delta_{x_t} \to 0$ under the usual bandwidth condition $h \to 0$ but $hn \to \infty$.*

Remark 3.2 As an example, suppose $g(x) = x$ and the L_2–optimal point predictor of $Y_{\rm f}$ is chosen in which case $\Pi \simeq m_{x_{\rm f}}$. Then, our $(1-\alpha)100\%$ equal-tailed, predictive interval for $Y_{\rm f}$ boils down to $[m_{x_{\rm f}} + q(\alpha/2),\ m_{x_{\rm f}} + q(1-\alpha/2)]$ where $q(\alpha)$ is the α—quantile of the empirical distribution of the B bootstrap root replicates of type $Y_{\rm f}^* - m_{x_{\rm f}}^*$.

Remark 3.3 As in all nonparametric smoothing problems, choosing the bandwidth is often a key issue due to the ever-looming problem of bias; the addition of a bootstrap algorithm as above further complicates things. In the closely related problem of constructing bootstrap confidence bands in nonparametric regression, different authors have used various tricks to account for the bias. For example, Härdle and Bowman (1988) construct a kernel estimate for the second derivative $\mu''(x)$, and use this estimate to explicitly correct for the bias; the estimate of the second derivative is known to be consistent but it is difficult to choose its bandwidth. Härdle and Marron (1991) estimate the (fitted) residuals using the optimal bandwidth but the resampled residuals are then added to an oversmoothed estimate of μ; they then smooth the bootstrapped data using the optimal bandwidth. Neumann and Polzehl (1998) use only one bandwidth but it is of smaller order than the mean square error optimal rate; this *undersmoothing* of curve estimates was first proposed by Hall (1993) and is perhaps the easiest theoretical solution towards confidence band construction although the recommended degree of undersmoothing for practical purposes is not obvious. In a recent paper, McMurry and Politis (2008) show that the use of infinite-order, flat-top kernels alleviates the bias problem significantly permitting the use of the optimal bandwidth. Although the above literature pertains to confidence intervals, the construction of prediction intervals is expected to suffer from similar difficulties; see Section 4.6. Furthermore, note the possible advantage of excluding residuals obtained from boundary points from the resampling procedure; see Remark 4.6 for more discussion.

Remark 3.4 An important feature of all bootstrap procedures is that they can handle *joint* prediction intervals, i.e., prediction *regions*, with the same ease as the univariate ones. For example, $x_{\rm f}$ can represent a collection of p 'future' x–points in the above Resampling Algorithm. The only difference is that in Step 2(c) we would need to draw p pseudo-errors r randomly (with replacement) from the set $(r_1, ..., r_n)$, and thus construct p bootstrap pseudo-responses, one for each of the p points in $x_{\rm f}$. Then, Step 5 of the Algorithm would give a multivariate (joint) predictive distribution for the response Y at the p points in $x_{\rm f}$ from which a joint prediction *region* can be extracted. If it is desired that the prediction region is of rectangular form, i.e., joint prediction *intervals* as opposed to a general-shaped region, then these can be based on the distribution of the maximum (and minimum) of the p targeted responses that is obtainable from the multivariate predictive distribution via the continuous mapping theorem.

For completeness, we now briefly discuss the predictive interval that follows from an assumption of normality of the errors ε_t in the model (8). In that case, $m_{x_{\rm f}}$ is also normal, and independent of the 'future' error $\varepsilon_{\rm f}$. If $\sigma^2(x)$ can be assumed to be at least as smooth as $\mu(x)$, then a normal approximation to the distribution of the root $Y_{\rm f} - m_{x_{\rm f}}$ implies an approximate $(1-\alpha)100\%$ equal-tailed, predictive interval for $Y_{\rm f}$ given by:

$$[m_{x_{\rm f}} + V_{x_{\rm f}} \cdot z(\alpha/2),\ m_{x_{\rm f}} + V_{x_{\rm f}} \cdot z(1-\alpha/2)] \tag{24}$$

where $V^2_{x_{\rm f}} = s^2_{x_{\rm f}} \left(1 + \sum_{i=1}^n \tilde{K}^2(\frac{x_{\rm f}-x_i}{h})\right)$ with $\tilde{K}$ defined in eq. (10), and $z(\alpha)$ being the α–quantile of the standard normal. If the 'density' (e.g. histogram) of the design points $x_1, ..., x_n$ can be thought to approximate a given functional shape (say, $f(\cdot)$) for large n, then the large-sample approximation

$$\sum_{i=1}^n \tilde{K}^2(\frac{x_{\rm f}-x_i}{h}) \sim \frac{\int K^2(x)dx}{nh\ f(x_{\rm f})} \tag{25}$$

can be used–provided $K(x)$ is such that $\int K(x)dx = 1$; see e.g. Li and Racine (2007).

Interval (24) is problematic in at least two respects: (a) it completely ignores the bias of m_x, so it must be either explicitly bias-corrected, or a suboptimal bandwidth must be used to ensure undersmoothing; and (b) it crucially hinges on *exact*, finite-sample normality of the data as its validity can not be justified by a central limit approximation. For all the above, the usefulness of interval (24) is quite limited.

3.6 Application: better prediction intervals in linear regression

The literature on predictive intervals in regression is not large; see e.g. Caroll and Ruppert (1991), Patel (1989), Schmoyer (1992) and the references therein. Furthermore, the liter-

ature on predictive distributions seems virtually non-existent outside the Bayesian framework. What is most striking is that even the problem of undercoverage of prediction prediction intervals in *linear* regression reported 25 years ago by Stine (1985) has not been satisfactorily resolved to this day; see the recent paper by Olive (2007).

Thus, in this subsection we focus on the usual linear regression model:

$$Y_i = \underline{x}_i' \underline{\beta} + Z_i, \text{ for } i = 1, \ldots, n, \tag{26}$$

with $Z_t \sim$ i.i.d. $(0,\sigma^2)$. Equivalently, $\underline{Y}_n = X\underline{\beta} + \underline{Z}_n$ where $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ and $\underline{Z}_n = (Z_1, \ldots, Z_n)'$ are $n \times 1$ random vectors, $\underline{\beta}$ is a $p \times 1$ deterministic parameter vector, and X is an $n \times p$ deterministic design matrix of full rank with ith row given by vector $\underline{x}_i'$.

Let $\hat{\underline{\beta}}$ be an estimator of $\underline{\beta}$ that is linear in the data $\underline{Y}_n$ so that the MF/MB methodology of Section 3.4, and in particular eq. (20), applies; an obvious possibility is the Least Squares (LS) estimator. Also let $\hat{\underline{\beta}}^{(i)}$ be the same estimator based on the delete-Y_i dataset. The predictive and fitted residuals ($\tilde{z}_i$ and z_i respectively) corresponding to data point Y_i are defined in the usual manner, i.e., $\tilde{z}_i = Y_i - \underline{x}_i' \hat{\underline{\beta}}^{(i)}$. and $z_i = Y_i - \underline{x}_i' \hat{\underline{\beta}}$. Analogously to eq. (21), here too the predictive residuals are always larger in absolute value (i.e., 'inflated') as compared to the fitted residuals. To see this, recall that

$$\tilde{z}_i = \frac{z_i}{1 - h_i}, \text{ for } i = 1, \ldots, n, \tag{27}$$

where $h_i = \underline{x}_i'(X'X)^{-1}\underline{x}_i$ is the ith diagonal element of the 'hat' matrix $X(X'X)^{-1}X'$; see e.g. Seber and Lee (2003, Th. 10.1), or Efron and Tibshirani (1993, ex. 17.1). Assuming that the regression has an intercept term, eq. (10.12) of Seber and Lee (2003) further implies $1/n \leq h_i \leq 1$ from which it follows that $|\tilde{z}_i| \geq |z_i|$ for all i.

Noting that the fitted residuals have variance depending on h_i, Stine (1985) suggested resampling the *studentized* residuals $\hat{z}_i = z_i/\sqrt{1 - h_i}$ in his construction of bootstrap prediction intervals. The studentized residuals $\hat{z}_i$ are also 'inflated' as compared to the fitted residuals z_i, so Stine's (1985) suggestion was an effort to reduce the undercoverage of bootstrap prediction intervals that was first pointed out by Efron (1983). However, Stine's proposal does not seem to fully correct the problem; for example, Olive (2007) recommends the use of an *ad hoc* further inflation of the residuals arguing that "since residuals underestimate the errors, finite sample correction factors are needed".

Nevertheless, it is apparent from the above discussion that $|\tilde{z}_i| \geq |\hat{z}_i|$. Hence, using the predictive residuals is not only intuitive and natural as motivated by the model-free prediction principle, but it also goes further towards the goal of increasing coverage without cumbersome (and arbitrary) correction factors.[7] To obtain predictive intervals for Y_{f}, the

[7]Efron (1983) proposed an iterated bootstrap method in order to correct the downward bias of the bootstrap estimate of prediction error; his method notably involved the use of predictive residuals albeit at the 2nd bootstrap tier—see Efron and Tibshirani (1993, Ch. 17.7) for details.

Resampling Algorithm of Section 3.5 now applies *verbatim* with the understanding that in the linear regression setting $m_x \equiv \underline{x}'\hat{\underline{\beta}}$.

As the following subsection confirms, the MF/MB method based on predictive residuals seems to correct the undercoverage of bootstrap prediction intervals. Finally, note that the methodology of Section 3.5 can equally address the *heteroscedastic* case when $Var(Z_i) = \sigma^2(\underline{x}_i)$, and an estimate of $\sigma^2(\underline{x}_i)$ is available via parametric or nonparametric methods.

3.7 Simulation: better prediction intervals in linear regression

We now conduct a small simulation in the linear regression setup of subsection 3.6 with $p = 2$, i.e., $\underline{x}_i = (1, x_i)'$, and $Y_i = \beta_0 + \beta_1 x_i + Z_i$, for $i = 1, \ldots, n$. For the simulation, the values $\beta_0 = -1$ and $\beta_1 = 1$ were used, and $Z_t \sim$ i.i.d. (0,1) from distribution Normal or Laplace. The design points $x_1, ..., x_n$ for $n = 50$ were generated from a standard normal distribution, and the prediction carried out at the point $x_{\rm f} = 1$. The simulation focused on constructing 90% prediction intervals, and was based on 900 repetitions of each experiment. Both LS regression and L_1 regression were considered for estimating β_0 and β_1.

Table 3.2 reports the empirical coverage levels (CVR), and (average) lower and upper limits of the different prediction intervals in the linear regression case. The standard error of the CVR entries is 0.01; the provided standard error (st.err.) applies equally to either the lower or upper limit of the interval. For the first five rows of Table 3.2, β_0 and β_1 were estimated by Least Squares which is optimal in the Normal case; in the last two rows of Table 3.2, β_0 and β_1 are estimated via L_1 regression which is optimal in the Laplace case. Note that the ideal point predictor of Y at $x_{\rm f} = 1$ is zero; so the prediction intervals are expected to be centered around zero. Indeed, all (average) intervals of Table 3.2 are approximately symmetric around zero.

Linear regression is, of course, a model-based setup; so both interval constructions MB (=model-based) and MF/MB (=model-free/model-based) of Section 3.5 are applicable; they were both considered here in addition to three competing intervals: Stine's (1985) interval that is analogous to the MB construction except that Stine used the studentized residuals; the usual NORMAL theory interval, namely $m_{x_{\rm f}} \pm t_{n-2}(\alpha/2)S\sqrt{1+h_{\rm f}}$; and Olive's (2007) 'semi-parametric' interval:

$$\left(m_{x_{\rm f}} + a_n e(\alpha/2)\sqrt{1+h_{\rm f}},\ m_{x_{\rm f}} + a_n e(1-\alpha/2)\sqrt{1+h_{\rm f}}\right).$$

In the above, $m_{x_{\rm f}}$ is the usual point predictor given by $\hat{\beta}_0 + \hat{\beta}_1 x_{\rm f}$, $h_{\rm f} = \underline{x}_{\rm f}'(X'X)^{-1}\underline{x}_{\rm f}$ is the 'leverage' at point $x_{\rm f}$, and $S^2 = (n-2)^{-1}\sum_{i=1}^n e_i^2$. In Olive's interval, $e(\alpha)$ is the α (sample) quantile of the residuals $\{e_1, ..., e_n\}$, and $a_n = (1 + \frac{15}{n})\sqrt{\frac{n}{n-2}}$ is an *ad hoc* 'correction' factor designed to increase coverage.

The findings of Table 3.2 are quite interesting:

- The NORMAL theory interval (based on t–quantiles) has exact coverage with Normal data—as expected—but slightly over-covers in the Laplace case. It is also the interval with smallest length variability.
- Olive's interval shows striking *over*-coverage which is an indication that the a_n correction factor is too extreme. Also surprising is the large variability in the length of Olive's interval that is 50% larger than that of our bootstrap methods.
- Looking at rows 1—3, the expected monotonicity in terms of increasing coverage is observed; i.e., CVR(MB) < CVR(MB Stine) < CVR(MF/MB).
- The MF/MB intervals have (almost) uniformly better coverage than their MB analogs indicating that using the predictive residuals is indeed the solution to the widely reported undercoverage of MB and Stine's intervals.

Distribution:	Normal			Laplace		
Case $x_f = 1$	CVR	INTERVAL	(st.err.)	CVR	INTERVAL	(st.err.)
MF/MB	0.890	$[-1.686, 1.682]$	(.011)	0.901	$[-1.685, 1.691]$	(.016)
MB	0.871	$[-1.631, 1.609]$	(.011)	0.886	$[-1.611, 1.619]$	(.015)
MB Stine	0.881	$[-1.656, 1.641]$	(.011)	0.892	$[-1.640, 1.663]$	(.015)
MB Olive	0.941	$[-2.111, 2.097]$	(.017)	0.930	$[-2.072, 2.089]$	(.025)
NORMAL	0.901	$[-1.723, 1.711]$	(.009)	0.910	$[-1.699, 1.716]$	(.011)
MF/MB L_1	0.896	$[-1.715, 1.709]$	(.012)	0.908	$[-1.699, 1.705]$	(.016)
MB L_1	0.871	$[-1.647, 1.632]$	(.012)	0.896	$[-1.619, 1.636]$	(.015)

Table 3.2. Empirical coverage levels (CVR), and (average) lower and upper bounds of different prediction intervals with nominal coverage of 0.90 in linear regression; the standard error (st.err.) applies equally to either the lower or upper limit.

4 Model-free prediction in regression

4.1 Constructing the transformation

We now revisit the nonparametric regression setup of Section 3 but in a situation where a model such as eq. (8) can not be considered to hold true (not even approximately). As an example of model (8) not being valid, consider the setup where the skewness and/or kurtosis of Y_t depends on x_t, and thus centering and studentization will not result in 'i.i.d.–ness'. For

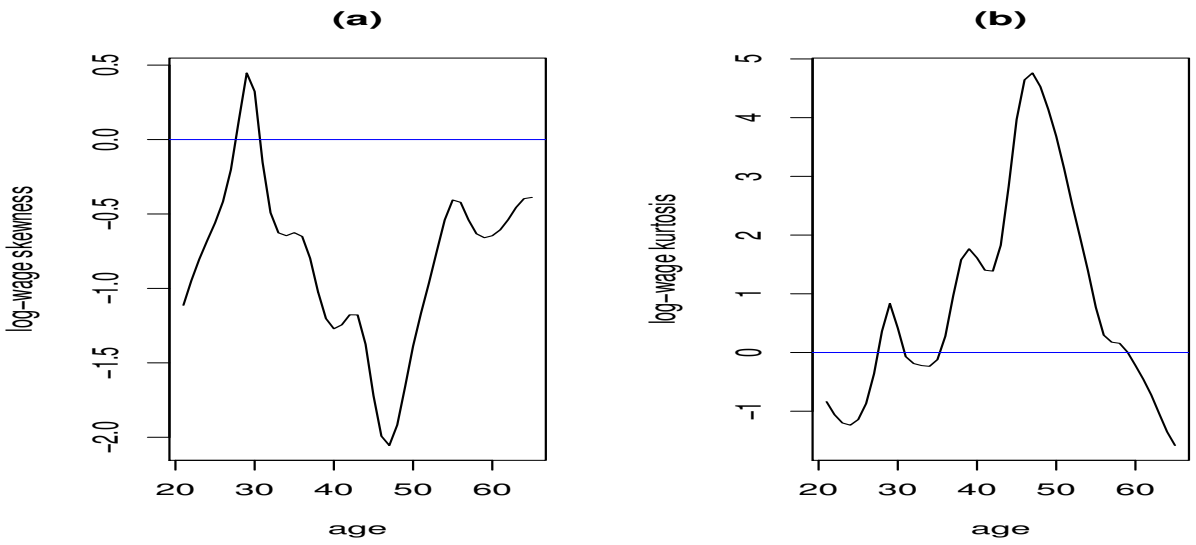

Figure 2: (a) Skewness of log-wage vs. age. (b) Kurtosis of log-wage vs. age. [Kernel-based estimates from dataset `cps71`.]

example, kernel estimates of skewness and kurtosis from dataset `cps71`—although slightly undersmoothed—clearly point to the non-constancy of these two functions; see Figure 2.

Throughout Section 4, the dataset is still $\{(Y_t, x_t),\ t = 1, \ldots, n\}$ where the regressor x_t is again assumed univariate and deterministic, and the Y_ts are independent although not identically distributed. We will denote their conditional distribution by

$$D_x(y) = P\{Y_{\mathrm{f}} \leq y | x_{\mathrm{f}} = x\}$$

where $(Y_{\mathrm{f}}, x_{\mathrm{f}})$ represents the random response Y_{f} associated with predictor x_{f}.

We will assume throughout that the quantity $D_x(y)$ is *continuous* in both x and y. To elaborate, we assume $D_x(y)$ to be continuous in y, i.e., that $Y_1, \ldots, Y_n$ are continuous random variables, since otherwise standard methods like Generalized Linear Models can be invoked, e.g. logistic regression, Poisson regression, etc.; see McCullagh and Nelder (1983). Furthermore, we assume that the collection of functions $D_x(\cdot)$ depends in a smooth way on x in order to make use of local regression ideas. Consequently, we can estimate $D_x(y)$ by a 'local' empirical distribution such as

$$N_{x,h}^{-1} \sum_{t:|x_t - x| < h/2} \mathbf{1}\{Y_t \leq y\} \tag{28}$$

where $\mathbf{1}\{Y_t \leq y\}$ denotes the indicator of event $\{Y_t \leq y\}$, and $N_{x,h}$ is the number of

summands, i.e., $N_{x,h} = \#\ \{t : |x_t - x| < h/2\}$. More generally, we can estimate $D_x(y)$ by

$$\hat{D}_x(y) = \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\} \tilde{K}\left(\frac{x - x_i}{h}\right) \tag{29}$$

where $\tilde{K}\left(\frac{x-x_i}{h}\right) = K\left(\frac{x-x_i}{h}\right) / \sum_{k=1}^{n} K\left(\frac{x-x_k}{h}\right)$ as before; for any fixed y, this is just a Nadaraya-Watson smoother of the variables $\mathbf{1}\{Y_t \leq y\},\ t = 1, \ldots, n$. Note that eq. (28) is just $\hat{D}_x(y)$ with K chosen as the rectangular kernel, i.e., $K(x) = \mathbf{1}\{|x| \leq 1/2\}$; in general, we can use any non-negative, integrable kernel $K(\cdot)$ in (29).

Remark 4.1 For $\hat{D}_x$ to be an accurate estimator of D_x, the value x must be such that it has an appreciable number of h-close neighbors among the original predictors $x_1, ..., x_n$, i.e., that the number $N_{x,h}$ is not too small. For example, if $N_{x,h} \leq 1$ the estimation of D_x is not just inaccurate—it is simply infeasible.

Estimator $\hat{D}_x(y)$ enjoys many good properties including asymptotic consistency under regularity conditions. For example,

$$Var(\hat{D}_x(y)) = O(\frac{1}{hn}) \ \text{ and } \ Bias(\hat{D}_x(y)) = O(h^2) \tag{30}$$

with $h \to 0$ but such that $hn \to \infty$; see Theorem 6.1 of Li and Racine (2007). Nevertheless, $\hat{D}_x(y)$ is discontinuous as a function of y, and therefore unacceptable for our purposes. In Politis (2010) a piecewise linear—and strictly increasing—version of $\hat{D}_x(y)$ was proposed; here, we will take a slightly different approach.

Observe that the discontinuity of $\hat{D}_x(y)$ as a function of y stems from the discontinuity of the indicator functions $\mathbf{1}\{Y_t \leq y\}$. We may therefore replace $\mathbf{1}\{Y_t \leq y\}$ by $\Lambda\left(\frac{y-Y_t}{b}\right)$ in eq. (29) leading to the estimator

$$\bar{D}_x(y) = \sum_{i=1}^{n} \Lambda\left(\frac{y - Y_t}{b}\right) \tilde{K}\left(\frac{x - x_i}{h}\right) \tag{31}$$

that is also studied in Li and Racine (2007, Ch. 6). In the above, b is a positive bandwidth parameter and $\Lambda(y)$ is a smooth distribution function that is strictly increasing, rendering the estimator $\bar{D}_x(y)$ continuous and strictly increasing in y.

For example, we may define $\Lambda(y) = \int_{-\infty}^{y} \lambda(s)ds$ where $\lambda(s)$ is a symmetric density function that is continuous and everywhere positive. In this case, it is apparent that $\bar{D}_x(y)$ will not only be continuous—it will actually be differentiable with respect to y. Thus, a different interpretation of estimator $\bar{D}_x(y)$ is that it is the indefinite integral of a local estimate of the *density* associated with distribution $D_x(y)$, i.e., an estimate of the derivative of $D_x(y)$ with respect to y (provided that exists).

Remark 4.2 Note that a local linear (or polynomial) smoother of the indicator variables $\mathbf{1}\{Y_t \leq y\}$ or the smooth variables $\Lambda\left(\frac{y-Y_t}{b}\right)$ could be used in place of the local constant estimators (29) and (31); this may actually be preferable in view of better handling of edge effects and non-equally spaced x–points. Details of these methods could be found in Li and Racine (2007) but the essence of our discussion here remains unchanged. Furthermore, the discussion of Remark 3.1 applies here as well, i.e., on possibly using a local bandwidth or a k–nearest neighbor smoother of $\mathbf{1}\{Y_t \leq y\}$ and/or $\Lambda\left(\frac{y-Y_t}{b}\right)$.

Fact 4.1 *Under regularity conditions—that include the use of an arbitrary (nonnegative) kernel $K(\cdot)$ and a regular "density" (i.e., histogram) of the design points $x_1, \ldots, x_n$ as required for eq.* (25)*), it follows that $\bar{D}_x(y)$ satisfies an equation similar to eq.* (30)*, namely:*

$$Var(\bar{D}_x(y)) = O(\frac{1}{hn}) \quad and \quad Bias(\bar{D}_x(y)) = O(h^2 + b^2) \tag{32}$$

assuming that $h \to 0$, $b \to 0$, $hn \to \infty$ and $\sqrt{hn}(h^3 + b^3) = o(1)$; see Theorem 6.2 of Li and Racine (2007). Thus, if $b = O(h)$, then estimator $\bar{D}_x(y)$ has Mean Squared Error (MSE) that is of the same asymptotic order as that of $\hat{D}_x(y)$. In order to minimize the asymptotic MSE of $\bar{D}_x(y)$, the optimal bandwidth specifications are $h \sim c_h n^{-1/5}$ and $b \sim c_b n^{-2/5}$ for some positive constants c_h, c_b.

Note that since $D_x(y)$ is assumed continuous (in y), the asymptotic consistency of $\bar{D}_x(y)$ implies that the inverse $\bar{D}_x^{-1}(\alpha)$ will also be consistent for $D_x^{-1}(\alpha)$ for any $\alpha \in [0, 1]$ as long as $D_x(y)$ is strictly increasing at $y = D_x^{-1}(\alpha)$; see e.g. Lemma 1.2.1 of Politis et al. (1999).

Recall that the Y_ts are non-i.i.d. only because they do not have identical distributions. Since they are continuous random variables, the *probability integral trasform* is the key idea to transform them towards 'i.i.d.–ness'. To see why, note that if we let

$$\eta_i = D_{x_i}(Y_i) \quad \text{for} \quad i = 1, \ldots, n$$

our transformation objective would be exactly achieved since $\eta_1, \ldots, \eta_n$ would be i.i.d. Uniform(0,1). Of course, $D_x(\cdot)$ is not known but we have the consistent estimator $\bar{D}_x(\cdot)$ as its proxy. Therefore, our proposed transformation amounts to defining

$$u_i = \bar{D}_{x_i}(Y_i) \quad \text{for} \quad i = 1, \ldots, n; \tag{33}$$

by the consistency of $\bar{D}_x(\cdot)$, we can now claim that $u_1, \ldots, u_n$ are approximately i.i.d. Uniform(0,1).

If a parametric specification for $D_x(y)$ happens to be available, i.e., if $P\{Y_t \leq y | x_t = x\}$ has known form up to a finite-dimensional parameter θ—that in general will depend on x—, then obviously our probability integral trasform of Y_t would be based on the parametric

distribution with parameter θ estimated from a local neighborhood of the associated regressor x_t. The probability integral trasform has been used by Ruppert and Cline (1994) as an intermediate step towards building better density estimators; however, our application is quite different as the following sections make clear.

4.2 Model-free optimal predictors

Since a transformation of the data towards 'i.i.d.–ness' is available from eq. (33), we can now formulate optimal predictors in the model-free paradigm. The key idea is to invert the probability integral trasform; to do this, we will be using the inverse transformation $\bar{D}_x^{-1}$ which is well-defined since $\bar{D}_x(\cdot)$ is strictly increasing by construction. Note that, for any $i = 1, ..., n$, $\bar{D}_{x_\mathrm{f}}^{-1}(u_i)$ is a *bona fide* potential response Y_f associated with predictor x_f since $\bar{D}_{x_\mathrm{f}}^{-1}(u_i)$ has (approximately) the same distribution as Y_f. These n valid potential responses given by $\{\bar{D}_{x_\mathrm{f}}^{-1}(u_i)$ for $i = 1, ..., n\}$ can be gathered together to give us an approximate empirical distribution for Y_f from which our predictors will be derived.

Thus, analogously with the discussion associated with the entries of Table 3.1 in Section 3, it follows that *the L_2—optimal predictor of $g(Y_\mathrm{f})$ will be the expected value of $g(Y_\mathrm{f})$ that is approximated by*

$$n^{-1} \sum_{i=1}^{n} g\left(\bar{D}_{x_\mathrm{f}}^{-1}(u_i)\right). \tag{34}$$

Similarly, the L_1—optimal predictor of $g(Y_\mathrm{f})$ will be approximated by the sample median of the set $\{g\left(\bar{D}_{x_\mathrm{f}}^{-1}(u_i)\right),\ i = 1, ..., n\}$. The model-free predictors are summarized in Table 4.1 that can be compared to Table 3.1 of the previous section.[8]

	Model-free (MF^2)
L_2—predictor of Y_f	mean$\{\bar{D}_{x_\mathrm{f}}^{-1}(u_i)\}$
L_1—predictor of Y_f	median$\{\bar{D}_{x_\mathrm{f}}^{-1}(u_i)\}$
L_2—predictor of $g(Y_\mathrm{f})$	mean$\{g\left(\bar{D}_{x_\mathrm{f}}^{-1}(u_i)\right)\}$
L_1—predictor of $g(Y_\mathrm{f})$	median$\{g\left(\bar{D}_{x_\mathrm{f}}^{-1}(u_i)\right)\}$

Table 4.1. The model-free (MF^2) optimal point predictors where $u_i = \bar{D}_{x_i}(Y_i)$.

Note that any of the two optimal model-free predictors (mean or median) can be used to give the equivalent of a model *fit*. To fix ideas, suppose we focus on the L_2—optimal

[8]For $\bar{D}_{x_\mathrm{f}}^{-1}$ to be an accurate estimator of $D_{x_\mathrm{f}}^{-1}$, the value x_f must be such that it has an appreciable number of h-close neighbors among the original predictors $x_1, ..., x_n$ as discussed in Remark 4.1. As an extreme example, note that prediction of Y_f when x_f is outside the range of the original predictors $x_1, ..., x_n$, i.e., extrapolation, is *not* feasible in the model-free paradigm.

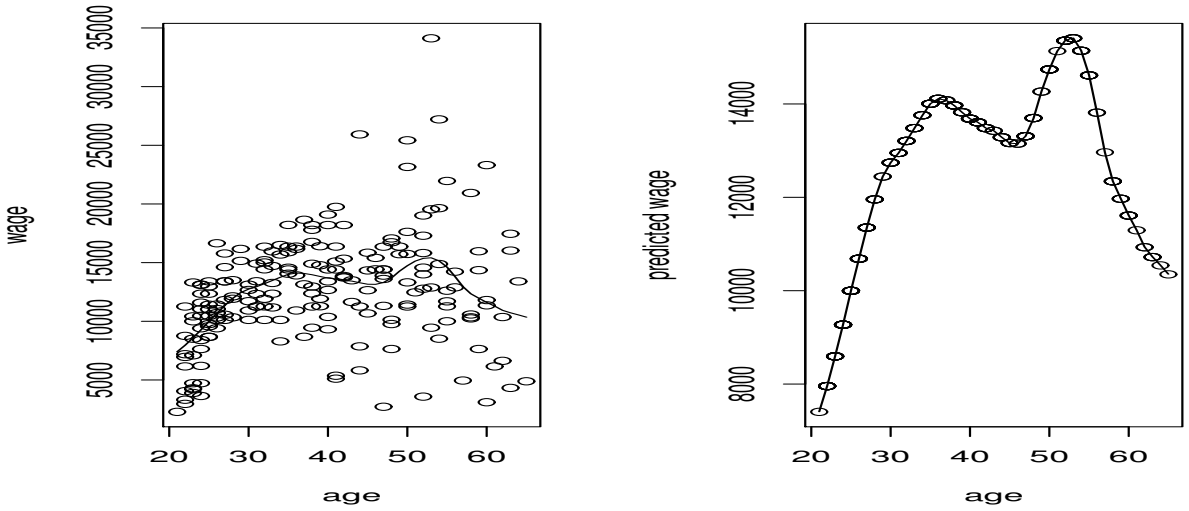

Figure 3: (a) Wage vs. age scatterplot. (b) Circles indicate the salary predictor from eq. (34) calculated from log-wage data with $g(x)$ exponential. For both figures, the superimposed solid line represents the MF^2 salary predictor calculated from the *raw* data (without the log-transformation).

case and that $g(x) = x$. Calculating the value of the optimal predictor of eq. (34) for many different $x_{\rm f}$ values, e.g., taken on a grid, the equivalent of a nonparametric smoother of a regression function is constructed, and can be plotted over the (Y, x) scatterplot. In this sense, *model-free model-fitting* (MF^2) is achieved as discussed in Remark 2.1.

Recall that the L_2—optimal predictor of $Y_{\rm f}$ associated with design point $x_{\rm f}$ is simply the conditional expectation $E(Y_{\rm f}|x_{\rm f})$. The latter is well approximated by our kernel estimator $m_{x_{\rm f}}$ (or a local polynomial) even *without* the validity of model (8), therefore also qualifying to be called a model-free *point* predictor. Predictor (34) can then be seen as an alternative method to estimate $E(Y_{\rm f}|x_{\rm f})$ which is actually close to $m_{x_{\rm f}}$ although not identical. To appreciate why, recall that if a random variable Y has distribution $F(y) = P(Y \le y)$, then $EY = \int y\, dF(y) = \int_0^1 F^{-1}(u)\, du$; it is the latter expression—in its local (in x) form—that predictor (34) approximates. Similarly, for the L_1—predictor we have: median$\{\bar{D}_{x_{\rm f}}^{-1}(u_i)\} = \bar{D}_{x_{\rm f}}^{-1}(\text{median}\{u_i\}) \simeq \bar{D}_{x_{\rm f}}^{-1}(1/2)$ since the u_is are approximately Uniform (0,1). Hence, the model-free L_1–optimal point predictor is close to the median of the (estimated) conditional distribution of $Y_{\rm f}$.

The real advantages of the model-free philosophy, however, are twofold: (a) it allows us to go *beyond* the point predictions and obtain valid predictive distributions and intervals for $Y_{\rm f}$ as will be described in Section 4.4—this is simply not possible on the basis of the

kernel estimator m_{x_f} without resort to a model like (8); and (b) it is a totally *automatic* method that does not require any preliminary preprocessing and/or data transformations; see Remark 4.3 below.

Remark 4.3 The model-free prediction technique based on transformation (33) is totally automatic, *relieving the practitioner from the need to find an optimal transformation* for additivity and variance stabilization; this is a significant practical advantage because of the multitude of such proposed transformations, e.g. the Box/Cox power family, ACE, AVAS, etc.—see Linton et al. (2008) and the references therein. For example, Figure 3 (a) is the analog of Figure 1 (a) using the raw salary data, i.e., without the logarithmic transformation. Superimposed is the MF^2 predictor of salary that uses transformation (33) on the raw data; as Figure 3 (b) shows, the latter is virtually identical to the MF^2 predictor obtained from the logarithmically transformed data and then using an exponential as the function $g(x)$ for predictor (34).

Remark 4.4 Following the discussion of Remark 4.2, it is now apparent that the model-free predictors of Table 4.1 are still computable in the case where the x–variable is discrete-valued provided, of course, that $N_{x,h}$ the number of data points in the local neighborhood of each of these discrete values is large enough to permit accurate estimation of $D_x(\cdot)$ locally. What allows the method to work here—and also to still work in terms of predictive intervals to be developed shortly—is that x_f will by necessity be one of these discrete values as well.

4.3 Cross-validation for model-free prediction

As seen in the last two subsections, estimating the conditional distribution $D_x(\cdot)$ by $\bar{D}_x(\cdot)$ is a crucial part of the model-free procedure; the accuracy of this estimation depends on the choice of bandwidth h. Recall that cross-validation is a predictive criterion since it aims at minimizing the sum of squares (or absolute values) of *predictive* residuals. Nevertheless, we can still from predictive residuals in model-free prediction, and thus cross-validation is possible in the model-free framework as well.

To fix ideas, suppose we focus on the L_2—optimal predictor of eq. (34), and let $\Pi_t^{(t)}$ denote the predictor of Y_t as computed from the delete-Y_t dataset: $\{(Y_i, x_i)$ for $i = 1, \ldots, t-1$ and $i = t+1, \ldots, n\}$, i.e., pretending the (Y_t, x_t) data pair is unavailable; this involves estimating $D_x(\cdot)$ by $\bar{D}_x^{(t)}(\cdot)$ computed from the delete-Y_t dataset, and having only $n-1$ values of u_i in connection with eq. (33) and (34). Finally, define the MF^2 predictive residuals

$$\tilde{e}_t = g(Y_t) - \Pi_t^{(t)} \quad \text{for} \quad t = 1, \ldots, n. \tag{35}$$

Choosing the best bandwidth h to use in our model-free predictor (34) can then be based on minimizing PRESS=$\sum_{t=1}^n \tilde{e}_t^2$ or PRESAR=$\sum_{t=1}^n |\tilde{e}_t|$ as before. If $\hat{D}_x$ and $\bar{D}_x$ are based

on k–nearest neighbor estimation as in Remark 4.2, then minimizing PRESS or PRESAR would yield the cross-validated choice of k to be used.

Note that cross-validation using the MF^2 predictive residuals of eq. (35) can be quite computationally expensive. In view of the discussion in the previous subsection argueing that the L_2—optimal predictor of eq. (34) is close to a kernel smoother of the $(g(Y), x)$ scatterplot, it follows that cross-validation on the latter should give a quick approximate solution to the bandwidth choice for the predictors of Table 4.1 as well.

4.4 Model-free predictive distributions and intervals

The empirical distribution of $g(Y_{\rm f})$ constructed in the Algorithm of Section 4.2 can not be regarded as a predictive distribution because it does not capture the variability of $\bar{D}_x$; resampling gives us a way out of this difficulty once again. Generally, the predictive distribution and prediction intervals for $g(Y_{\rm f})$ can be obtained by the resampling algorithm of Section 2.6 that is re-cast below in the model-free regression framework.

Let $g(Y_{\rm f}) - \Pi$ be the prediction root where Π is either the L_2– or L_1–optimal predictor from Table 4.1, namely $\Pi = n^{-1}\sum_{i=1}^{n} g\left(\bar{D}_{x_{\rm f}}^{-1}(u_i)\right)$ or $\Pi = \text{median}\,\{g\left(\bar{D}_{x_{\rm f}}^{-1}(u_i)\right)\}$. Then, our algorithm for MF^2 prediction intervals reads as follows.

Resampling Algorithm for MF^2 predictive distribution of $g(Y_{\rm f})$

1. Based on the Y–data, estimate the conditional distribution $D_x(\cdot)$ by $\bar{D}_x(\cdot)$, and use eq. (33) to obtain the transformed data $u_1, ..., u_n$ that are approximately i.i.d.

   (a) Sample randomly (with replacement) the transformed data $u_1, ..., u_n$ to create bootstrap pseudo-data $u_1^*, ..., u_n^*$ whose empirical distribution is denoted $\hat{F}_n^*$.

   (b) Use the inverse transformation $\bar{D}_x^{-1}$ to create pseudo-data in the Y domain, i.e., let $\underline{Y}_n^\star = (Y_1^\star, ..., Y_n^\star)$ where $Y_t^* = \bar{D}_{x_t}^{-1}(u_t^*)$.

   (c) Generate a bootstrap pseudo-response $Y_{\rm f}^*$ by letting $Y_{\rm f}^* = \bar{D}_{x_{\rm f}}^{-1}(u)$ where u is drawn randomly from the set $(u_1, ..., u_n)$.

   (d) Based on the pseudo-data $\underline{Y}_n^\star$, re-estimate the conditional distribution $D_x(\cdot)$; denote the bootstrap estimator by $\bar{D}_x^*(\cdot)$.

   (e) Calculate a replicate of the bootstrap root $g(Y_{\rm f}^*) - \Pi^*$ where $\Pi^* = n^{-1}\sum_{i=1}^{n} g\left(\bar{D}_{x_{\rm f}}^{*^{-1}}(u_i^*)\right)$ or $\Pi^* = \text{median}\,\{g\left(\bar{D}_{x_{\rm f}}^{*^{-1}}(u_i^*)\right)\}$ according to whether L_2– or L_1–optimal prediction has been used for the original Π.

2. Steps (a)—(e) in the above are repeated B times, and the B bootstrap root replicates are collected in the form of an empirical distribution with α–quantile denoted by $q(\alpha)$.

3. Then, the model-free $(1-\alpha)100\%$ equal-tailed, prediction interval for $g(Y_{\rm f})$ is

$$[\Pi + q(\alpha/2), \Pi + q(1-\alpha/2)] \tag{36}$$

and our estimate of the predictive distribution of $g(Y_{\rm f})$ is the empirical distribution of bootstrap roots obtained in step 2 shifted to the right by the number Π.

Remark 4.5 By construction, the $u_1, ..., u_n$ are aproximately i.i.d. $U(0,1)$; hence, steps 1(a) and 1(c) of the above algorithm could be modified to read:

(a′) Generate bootstrap pseudo-data $u_1^*, ..., u_n^*$ i.i.d. Uniform $(0,1)$, and denote $\hat{F}_n^*$ the empirical distribution of $u_1^*, ..., u_n^*$.

(c′) Generate a bootstrap pseudo-response $Y_{\rm f}^*$ by letting $Y_{\rm f}^* = \bar{D}_{x_{\rm f}}^{-1}(u)$ where u is drawn randomly from a Uniform $(0,1)$ distribution.

If the above choice is made, then there is no need to *'uniformize'* our data, and the step function $\hat{D}_x(\cdot)$ suffices as an estimator. As pointed out by a referee, one could then use the quantile inverse of $\hat{D}_x(\cdot)$ in place of the regular inverse $\bar{D}_x^{-1}(\cdot)$ wherever the latter is required; details (and comparisons) will be published elsewhere due to lack of space. For example, if an inspection of the transformed responses $u_1, \ldots, u_n$ shows that their distribution is *not* close to uniform, then resampling from the Uniform $(0,1)$ distribution may well be preferable. In any case, an inspection of $u_1, \ldots, u_n$ can be a very helpful diagnostic in much the same way as the usual residual diagnostics in regression—see Section 4.6 of Politis (2010) for more details.

Remark 4.6 Smoothing techniques are often plagued by edge effects, and this is especially true for kernel smoothers; Figures 1(a) and 3(a) show the bias problems near the left boundary. Thus, to implement the Resampling Algorithm for prediction intervals of this Section—but also to construct the point predictors of Table 4.1—it is practically advisable to only include the u_is obtained from x_is that are away from either boundary by more than half a bandwidth. From these u_is, a full-size resample $(u_1^*, ..., u_n^*)$ can be generated that, in turn, gives rise to a full-size pseudo-sample $(Y_1^*, ..., Y_n^*)$ which allows us to compute the bootstrap estimator $\bar{D}_x^*(\cdot)$. Similarly, only the Y^*s that are away from the boundaries by more than half a bandwidth will be used in the construction of Π^* in Step 1(e) above.

As a referee pointed out, the above recommendation has some caveats. For example, with data that are censored or truncated, there could be a boundary point carrying a cluster of data; it is apparent that this cluster is very informative and should not be excluded from the analysis (nor the bootstrap procedure). Recall that in Remark 4.4 it was put forth that it is possible to consider regression data associated with a regressor x that is discrete. The

censored/truncated data setup would be akin to regression data associated with a regressor x that is mixed (continuous and discrete); this does not pose a problem to the Model-free methodology. A qualitatively similar situation may occur when the density of the design points $x_1, \ldots, x_n$ is very skewed towards one of the end points; here, a simple solution might be to use a local bandwidth as previously mentioned.

4.5 Better model-free prediction intervals: MF/MF^2

The success of the MF/MB method of Section 3.5 is based on the fact that the distribution of the prediction error can be approximated better by the (empirical) distribution of the predictive residuals as compared to the (empirical) distribution of the fitted residuals. Using the latter—as in the MB method—typically results in variance underestimation and undercoverage of prediction intervals.

Since MF^2 predictive residuals are computable from eq. (35), one might be tempted to try to use them in order to mimic the MF/MB construction. Unfortunately, the MF^2 predictive residuals of eq. (35) are *not* i.i.d. in the context of the present section; hence, i.i.d. bootstrap on them is not recommended. In what follows, we will try to identify analogs of the i.i.d. predictive residuals in this model-free setting.

Recall that the accuracy of our bootstrap prediction intervals hinges on the accuracy of the approximation of the prediction root $g(Y_{\rm f}) - \Pi$ by its bootstrap analog, namely $g(Y_{\rm f}^*) - \Pi^*$. However, Π is based on a sample of size n, and $Y_{\rm f}$ is *not* part of the sample. Using predictive residuals is a trick that helps the bootstrap root mimic this situation by making $Y_{\rm f}^*$ into a genuinely "out-of-the-sample" point; the reason is that *every* data point is treated as an "out-of-the-sample" point as far as the computation of predictive residuals is concerned.

We can still achieve this effect within the MF^2 paradigm using an analogous trick; to see how, let $\bar{D}_{x_t}^{(t)}$ denote the estimator $\bar{D}_{x_t}$ as computed from the delete-Y_t dataset: $\{(Y_i, x_i)$, $i = 1, \ldots, t-1$ and $i = t+1, \ldots, n\}$. Now let

$$u_t^{(t)} = \bar{D}_{x_t}^{(t)}(Y_t) \quad \text{for} \quad t = 1, \ldots, n; \tag{37}$$

the $u_t^{(t)}$ variables will serve as the analogs of the predictive residuals $\tilde{e}_t$ of Section 3.5. Although the latter are approximately i.i.d. *only* when model (8) holds true, the $u_t^{(t)}$s are approximately i.i.d. in general under the weak assumptions of smoothness of $D_x(y)$.

Resampling Algorithm for MF/MF^2 predictive distribution of $g(Y_{\rm f})$

- The MF/MF^2 Resampling Algorithm is identical to the Algorithm for MF^2 predictive distribution of Section 4.4 with the following exception: replace the variables $u_1, ..., u_n$ by $u_1^{(1)}, ..., u_n^{(n)}$ throughout the construction.

The above Resampling Algorithm is denoted by MF/MF^2 to differentiate it from the algorithm of the previous subsection. The MF/MF^2 name alludes to the MF/MB construction of Section 3.5 to which it (approximately) reduces when model (8) happens to be true. Finally, the MF/MF^2 optimal point predictors are identical to the MF^2 predictors of Table 4.1 with the same exception: replace the variables $u_1, ..., u_n$ by $u_1^{(1)}, ..., u_n^{(n)}$.

4.6 Simulation: when a nonparametric regression model is true

The building block for the simulation in this subsection is model (8) with $\mu(x) = \sin(x)$, $\sigma(x) = 1/2$, and errors ε_t i.i.d. N(0,1) or two-sided exponential (Laplace) rescaled to unit variance. Knowledge that the variance $\sigma(x)$ is constant was not used in the estimation procedures, i.e., $\sigma(x)$ was estimated from the data. For each distribution, 500 datasets each of size $n = 100$ were created with the design points $x_1, \ldots, x_n$ being equi-spaced on $(0, 2\pi)$, and Nadaraya-Watson estimates of $\mu(x) = E(Y|x)$ and $\sigma^2(x) = Var(Y|x)$ were computed using a normal kernel in R.

Prediction intervals with nominal level $\alpha = 0.90$ were constructed using the two methods presented in Section 3: Model-Based (MB) and Model-Free/Model-Based (MF/MB); the two methods presented in Section 4: Model-Free (MF^2) and MF/MF^2; and the NORMAL approximation interval (24). For all methods (except the NORMAL) the correction of Remark 4.6 was employed. The required bandwidths were computed by L_1 (PRESAR) cross-validation. For simplicity—and to guarantee that $M_x \geq m_x^2$—equal bandwidths were used for both m_x and M_x, i.e., the constraint $h = q$ was imposed.

Before evaluating the performance of the resulting prediction intervals, it is of interest to check whether the u_i defined in (33) are indeed "uniformized" as their usage in the MF^2 and MF/MF^2 procedures requires. From each of the 500 replications, the set of $u_1, \ldots, u_n$ was constructed, and compared to the Uniform(0,1) via a Kolmogorov-Smirnov (K-S) test. Only 1 out of the 500 cases resulted in a rejection of the Uniform(0,1) null hypothesis at level 0.05. This could be regarded as good news for the "uniformize" procedure of eq. (33) but it also underscores an interesting issue: the variability of the K-S distances is smaller than that expected from i.i.d. Uniform(0,1) samples, and that is why the number of rejections is smaller than expected. The reason for this reduced variability could be attributed to the fact that the $u_1, \ldots, u_n$ are not exactly independent in our finite-sample setup; instead, they exhibit lag-1 and lag-2 autocorrelations of the order of -0.07 which is not statistically significant but nevertheless present. The negative—albeit small—autocorrelation may result into a reduced probability of clustering of the $u_1, \ldots, u_n$ data, and therefore explain the reduced variabiliy of the K-S statistics. Note, however, that this is a finite-sample effect; with a larger n, the bandwidth h decreases, and so does the correlation present in the $u_1, \ldots, u_n$ data. In any case, this correlation is destroyed in the bootstrap reshuffling that

is implemented in the MF^2 and MF/MF^2 procedures.

For each type of prediction interval constructed, the corresponding empirical coverage level (CVR) and average length (LEN) were recorded together with the (empirical) standard error associated with each average length. The standard error of the reported coverage levels over the 500 replications is 0.013; notably, these coverage levels represent *overall* (i.e., unconditional) probabilities in the terminology of Beran (1990); see also Cox (1975).

As previously mentioned, in the practical construction of bootstrap predictive intervals one would employ a large number of bootstrap simulations, say $B = 999$. Nevertheless, bootstrap predictive intervals are very computer-intensive; hence, for the purposes of our simulation this number was curtailed to $B = 249$. Even with $B = 249$ and with the generation of just 500 series for each scenario, the compilation of the entries of Table 4.2 took seven days of CPU time on a 2.5GHz PC. The R functions used in the computation are provided (with absolutely no warranty!) at: `http://www.math.ucsd.edu/~politis/SOFT/MF3functions.R`.

Tables 4.2 and 4.3 summarize our findings, and contain a number of important features:

- As mentioned before, the standard error of the reported CVRs is 0.013. In addition, note that—by construction—this simulation problem has some symmetry that helps us further appreciate the variability of the CVRs. For example, the expected CVRs should be the same for $x_f = 0.3\pi$ and 1.7π in all methods; so for the NORMAL case of Table 4.2(a), the CVR would be better estimated by the average of 0.876 and 0.859, i.e., closer to 0.868.

- The NORMAL intervals are characterized by under-coverage even when the true distribution is Normal. This under-coverage is a bit more pronounced when $x_f = \pi/2$ or $3\pi/2$ due to the high bias of the kernel estimator at the points of a 'peak' or 'valley' that the normal interval (24) 'sweeps under the carpet'.

- The length of the NORMAL intervals is quite less variable than those based on bootstrap; this should come as no surprise since the extra randomization implicit in any bootstrap procedure is expected to inflate the overall variances. [Note that the standard deviation of the length can be estimated by st. err. $\times\sqrt{500.}$]

- The MF/MB intervals are *always* more accurate (in terms of coverage) than their MB analogs in Tables 4.2(a) and 4.3(a). This was not unexpected since (i) the regression model (8) holds true here; (ii) bootstrap model-based intervals are expected to undercover; and (iii) by Fact 3.1, MF/MB intervals are expected to be wider, and therefore partially correct this under-coverage.

- The performance of MF^2 intervals shows a striking resemblance to that of MB intervals despite the fact that the former are constructed without making use of eq. (8);

similarly, the performance of MF/MF^2 intervals resembles that of MF/MB intervals.

- The overall winners in terms of coverage accuracy in Tables 4.2(a) and 4.3(a) appear to be the MF/MB and MF/MF^2 intervals. Nevertheless, both methods result in over-coverage when $x_{\mathrm{f}} \approx \pi$; this is due to the general phenomenon of *'bias leakage'* that will be discussed in more detail below.

- Although the coverage of the MF/MB and MF/MF^2 intervals is similar, the latter appear to have shorter average length (and smaller variability of interval length), and are thus preferable. This is quite surprising since one would expect that there would be a price to pay for using the more generally valid MF/MF^2 intervals instead of the model-specific MF/MB ones.

$x_{\mathrm{f}}/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	0.845	0.832	0.822	0.838	0.847	0.865	0.798	0.852	0.854
MF/MB	0.901	0.912	0.874	0.897	0.921	0.908	0.861	0.903	0.914
MF^2	0.834	0.836	0.829	0.831	0.849	0.852	0.804	0.840	0.838
MF/MF^2	0.897	0.906	0.886	0.886	0.912	0.895	0.868	0.897	0.899
Normal	0.874	0.876	0.872	0.867	0.863	0.877	0.865	0.870	0.868

Table 4.2(a). Empirical coverage levels (CVR) of prediction intervals according to different methods at several x_{f} points spanning the interval $(0, 2\pi)$. Nominal coverage was 0.90, sample size $n = 100$ and bandwidths chosen by L_1 cross-validation. Error distribution: i.i.d. Normal.

$x_{\mathrm{f}}/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	1.522	1.495	1.422	1.544	1.691	1.575	1.437	1.500	1.557
	0.013	0.013	0.012	0.014	0.014	0.013	0.012	0.013	0.013
MF/MB	1.824	1.785	1.699	1.844	2.013	1.895	1.715	1.798	1.856
	0.017	0.017	0.015	0.017	0.017	0.018	0.016	0.017	0.017
MF^2	1.501	1.467	1.430	1.507	1.621	1.537	1.441	1.481	1.519
	0.012	0.011	0.011	0.012	0.012	0.012	0.011	0.011	0.012
MF/MF^2	1.798	1.768	1.707	1.800	1.926	1.828	1.715	1.758	1.815
	0.013	0.013	0.012	0.013	0.013	0.012	0.012	0.012	0.013
Normal	1.595	1.591	1.591	1.593	1.591	1.592	1.591	1.591	1.594
	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005

Table 4.2(b). (Average) lengths (LEN)—with standard errors below them—of the prediction intervals reported in Table 4.2(a).

The problematic case $x_f \approx \pi$ deserves special discussion. In principle, this should be an easy case since kernel smoothers have approximately zero bias there. Nevertheless, smoothers will have appreciable bias at *all* other points where the curvature is nonzero, and in particular, at the peak/valley points $x_f = \pi/2$ and $x_f = 3\pi/2$. This bias is passed on to the residuals (fitted, predictive, or even the u_i variables of MF^2 and MF/MF^2) in the following way: residuals obtained near the point $x_f = \pi/2$ will tend to be larger (their distribution being skewed right), while residuals near the point $x_f = 3\pi/2$ will tend to be smaller (more negative, i.e., skewed left). By the bootstrap reshuffling of residuals, the skewness disappears but an artificial inflation of the residual distribution ensues; this contamination of the residual pool may adversely influence the prediction interval coverage. This is the phenomenon previously referred to as *'bias leakage'* that is expected to result in *over*coverage of bootstrap prediction (or confidence) intervals at points where the regression function has small curvature. 'Bias leakage' would be alleviated with a larger sample size and/or using higher-order smoothing kernels or other low bias approximation methods, e.g., wavelets. It could also be alleviated using bandwidth tricks such as *undersmoothing*—see the detailed discussion in Remark 3.3. A different way out of this difficulty may be to use a version of *local* resampling as in Shi (1991); we will not pursue this further here due to lack of space.

$x_f/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	0.870	0.820	0.841	0.894	0.883	0.872	0.831	0.867	0.868
MF/MB	0.899	0.870	0.886	0.910	0.912	0.905	0.868	0.905	0.888
MF^2	0.868	0.805	0.834	0.867	0.874	0.859	0.825	0.858	0.856
MF/MF^2	0.905	0.876	0.885	0.916	0.910	0.910	0.865	0.905	0.895
Normal	0.884	0.890	0.872	0.875	0.889	0.889	0.874	0.888	0.887

Table 4.3(a). Empirical coverage levels (CVR) of prediction intervals according to different methods at several x_f points spanning the interval $(0, 2\pi)$. Nominal coverage was 0.90, sample size $n = 100$ and bandwidths chosen by L_1 cross-validation. Error distribution: i.i.d. Laplace.

$x_f/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	1.572	1.522	1.420	1.584	1.708	1.564	1.437	1.536	1.576
	0.017	0.015	0.015	0.016	0.018	0.016	0.017	0.017	0.017
MF/MB	1.855	1.787	1.674	1.873	2.025	1.825	1.682	1.804	1.874
	0.023	0.020	0.018	0.021	0.021	0.019	0.021	0.022	0.021
MF^2	1.513	1.471	1.390	1.518	1.613	1.515	1.408	1.483	1.515
	0.014	0.013	0.013	0.014	0.015	0.014	0.013	0.015	0.014
MF/MF^2	1.806	1.785	1.685	1.837	1.942	1.806	1.708	1.779	1.820
	0.018	0.018	0.017	0.018	0.019	0.018	0.018	0.019	0.018
Normal	1.591	1.589	1.589	1.589	1.589	1.589	1.589	1.589	1.590
	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006

Table 4.3(b). (Average) lengths (LEN)—with standard errors below them—of the prediction intervals reported in Table 4.3(a).

4.7 Simulation: when a nonparametric regression model is *not* true

In this subsection, we investigate the performance of the different prediction intervals in a setup where model (8) is not true. For easy comparison with Section 4.6, we will keep the same (conditional) mean and variance, i.e., we will generate independent Y data such that $E(Y|x) = \sin(x)$, $Var(Y|x) = 1/2$, and design points $x_1, \ldots, x_{100}$ equi-spaced on $(0, 2\pi)$ as before. However, the error structure $\varepsilon_x = (Y - E(Y|x))/\sqrt{Var(Y|x)}$ may have skewness and/or kurtosis that depends on x, thereby violating the i.i.d. assumption.

For our simulation we considered the simple construction:

$$\varepsilon_x = \frac{c_x Z + (1 - c_x)W}{\sqrt{c_x^2 + (1 - c_x)^2}} \tag{38}$$

where $c_x = x/(2\pi)$ for $x \in [0, 2\pi]$, and $Z \sim N(0, 1)$ independent of W that has mean zero and variance one but will have either an exponential shape, i.e., $\frac{1}{2}\chi^2_2 - 1$, to capture a changing *skewness*, or Student's t with 5 d.f., i.e., $\sqrt{\frac{3}{5}}\, t_5$, to capture a changing *kurtosis*.

$x_f/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	0.886	0.893	0.840	0.840	0.888	0.816	0.780	0.827	0.816
MF/MB	0.917	0.927	0.877	0.917	0.928	0.892	0.849	0.888	0.885
MF^2	0.894	0.876	0.823	0.843	0.876	0.813	0.782	0.816	0.831
MF/MF^2	0.917	0.932	0.897	0.894	0.930	0.874	0.836	0.881	0.890
Normal	0.917	0.911	0.896	0.891	0.890	0.884	0.870	0.863	0.848

Table 4.4(a). Empirical coverage levels (CVR) of prediction intervals according to different methods at several x_f points spanning the interval $(0, 2\pi)$. Nominal coverage was 0.90, sample size $n = 100$ and bandwidths chosen by L_1 cross-validation. Error distribution: non–i.i.d. skewed as in eq. (38) with χ^2_2.

$x_f/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	1.482	1.460	1.369	1.529	1.700	1.543	1.398	1.484	1.543
	0.020	0.018	0.016	0.016	0.016	0.014	0.012	0.013	0.013
MF/MB	1.756	1.750	1.617	1.825	2.032	1.829	1.678	1.779	1.834
	0.024	0.024	0.020	0.022	0.022	0.017	0.020	0.020	0.022
MF^2	1.424	1.413	1.359	1.454	1.599	1.498	1.395	1.457	1.494
	0.015	0.015	0.014	0.014	0.014	0.013	0.011	0.011	0.012
MF/MF^2	1.723	1.691	1.618	1.735	1.893	1.772	1.636	1.710	1.756
	0.019	0.018	0.016	0.017	0.016	0.014	0.013	0.013	0.013
Normal	1.602	1.599	1.599	1.599	1.599	1.600	1.599	1.599	1.601
	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006

Table 4.4(b). (Average) lengths (LEN)—with standard errors below them—of the prediction intervals reported in Table 4.4(a).

$x_f/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	0.836	0.849	0.804	0.868	0.872	0.831	0.805	0.840	0.858
MF/MB	0.883	0.885	0.850	0.921	0.923	0.908	0.861	0.890	0.921
MF^2	0.814	0.843	0.798	0.861	0.863	0.829	0.816	0.831	0.865
MF/MF^2	0.876	0.899	0.858	0.923	0.921	0.894	0.877	0.874	0.914
Normal	0.898	0.879	0.885	0.878	0.872	0.885	0.875	0.868	0.885

Table 4.5(a). Empirical coverage levels (CVR) of prediction intervals according to different methods at several x_f points spanning the interval $(0, 2\pi)$. Nominal coverage was 0.90,

sample size $n = 100$ and bandwidths chosen by L_1 cross-validation. Error distribution: non–i.i.d. kurtotic as in eq. (38) with t_5–distribution.

$x_f/\pi =$	0.15	0.3	0.5	0.75	1	1.25	1.5	1.7	1.85
MB	1.529	1.484	1.404	1.558	1.698	1.563	1.415	1.506	1.551
	0.017	0.015	0.015	0.015	0.015	0.013	0.012	0.012	0.013
MF/MB	1.814	1.743	1.656	1.850	1.997	1.851	1.688	1.794	1.837
	0.021	0.018	0.020	0.019	0.017	0.016	0.018	0.018	0.019
MF^2	1.474	1.436	1.375	1.497	1.615	1.532	1.420	1.486	1.523
	0.013	0.012	0.012	0.013	0.013	0.012	0.010	0.011	0.012
MF/MF^2	1.764	1.706	1.653	1.794	1.920	1.808	1.687	1.758	1.798
	0.016	0.014	0.015	0.015	0.014	0.013	0.012	0.013	0.013
Normal	1.601	1.598	1.598	1.599	1.598	1.599	1.598	1.598	1.601
	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006

Table 4.5(b). (Average) lengths (LEN)—with standard errors below them—of the prediction intervals reported in Table 4.5(a).

Tables 4.4 and 4.5 present our findings; they are qualitatively similar to those of Tables 4.2 and 4.3 although the problem at hand is more complicated because of the skewness or kurtosis changing with x. In particular:

- In Table 4.4(a), note the coverage of the NORMAL intervals decreases monotonically as x_f increases, yielding over-coverage in the region where skewness exists, and under-coverage in the region with (close to) normal errors. Such a clear pattern does not exist in Table 4.5(a), suggesting that the NORMAL intervals may be more influenced by data skewness rather than kurtosis (at least when the deviations in the latter are small); this is consistent with the Edgeworth expansion approach to the related problem of confidence interval construction.

- The similarity of the MB to MF^2 intervals, and of the MF/MB to MF/MF^2 intervals remains striking. The MB and MF^2 intervals exhibit general under-coverage, while the MF/MB and MF/MF^2 intervals have the best coverage overall. The MF/MF^2 intervals are still preferable due to their shorter average length (and smaller variability of interval length).

As a final note, recall that the MF^2 and MF/MF^2 rows of Tables 4.2–4.5 were constructed using the algorithms of Sections 4.4 and 4.5 respectively. The option of using the explicit Uniform (0,1) distribution as discussed in Remark 4.5 was also explored by simulation. The

results were qualitatively similar to the ones in Tables 4.2–4.5, and are thus not presented here to save space.

Conclusions

Prediction has been traditionally approached in a model-based fashion. In this paper, we outline a model-free approach to prediction based on a new *'model-free prediction principle'*. The idea behind this principle is transforming the data into a domain that is easier to work with, e.g., an i.i.d. and/or a Gaussian setup. As demonstrated in Sections 3 and 4, the model-free prediction principle works very well in the context of regression data.

In particular, model-free model-fitting yields intuitive point predictors that are very close to the corresponding model-based ones when a model is true without explicit resort to a model equation; see Tables 3.1 and 4.1 for a summary. In addition, it is shown how resampling ideas can be coupled with the MF^2 methodology in order to construct *frequentist* predictive distributions and intervals that are generally valid in the presence or absence of an additive regression model. As an aside, MF^2 gives an intuitive solution to the well-documented problem of under-coverage of bootstrap prediction intervals in linear regression without the need for *ad hoc* correction factors.

The model-free prediction principle suggests the way to do nonparametric regression when an additive model is not available (MF^2), as well as suggesting an improvement (MF/MB) when such a model is available. As a surprising by-product, the MF^2 methodology seems to obviate the need to search for optimal transformations in regression. Finite-sample simulations confirm the good performance of these prediction intervals, and compare the different variations. Potential problems and diagnostics regarding MF^2 implementation are discussed in Section 4.6 of Politis (2010).

References

[1] Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression, *American Statistician*, vol. 46, no. 3, 175-185.

[2] Atkinson, A.C.(1985). *Plots, Transformations and Regression*, Clarendon Press, Oxford.

[3] Beran, R. (1990). Calibrating prediction regions, *J. Amer. Statist. Assoc.*, 85, 715–723.

[4] Bickel, P. and Li, B. (2006). Regularization in Statistics, *Test*, vol. 15, no. 2, 271-344.

[5] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations, *J. R. Statist. Soc., Ser. B*, 26, 211-252.

[6] Breiman, L. and Friedman, J. (1985), Estimating optimal transformations for multiple regression and correlation, *J. Amer. Statist. Assoc.*, 80, 580-597.

[7] Carmack, P.S., Schucany, W.R., Spence, J.S., Gunst, R.F., Lin, Q., and Haley, R.W. (2009). Far casting cross-validation. *J. Comput. Graph. Statist.*, vol. 18, no. 4, 879–893.

[8] Carroll, R.J. and Ruppert, D. (1988). *Transformations and Weighting in Regression*, Chapman and Hall, New York.

[9] Cox, D.R. (1975). Prediction intervals and empirical Bayes confidence intervals, in *Perspectives in Probability and Statistics*, (J. Gani, Ed.), Academic Press, London, pp. 47-55.

[10] Dai, J. and Sperlich, S. (2010). Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation, *Comp. Statist. Data Anal.*, vol. 54, no. 11, pp. 2487–2497.

[11] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York.

[12] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*, Cambridge Univ. Press.

[13] Dawid, A.P. (2004). Probability, causality, and the empirical world: a Bayes–de Finetti–Popper–Borel synthesis, *Statist. Sci.,* vol. 19, no. 1, 44–57.

[14] Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis, 3rd Ed.,* Wiley, New York.

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

[16] Efron, B. (1983), Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78, 316-331.

[17] Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[18] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman and Hall, London.

[19] Freedman, D.A. (1981). Bootstrapping regression models, *Annals of Statistics,* 9, 1218-1228.

[20] Gangopadhyay, A.K. and Sen, P.K. (1990). Bootstrap confidence intervals for conditional quantile functions, *Sankhya, Ser. A.*, vol. 52, no. 3, pp. 346-363.

[21] Goldberger, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model, *J. Amer. Statist. Assoc.*, 57, 369-375.

[22] Geisser, S.(1993). *Predictive Inference: An Introduction*, Chapman and Hall, New York.

[23] Hahn, J. (1995). Bootstrapping quantile regression estimators, *Econometric Theory*, vol. 11, no. 1, pp. 105-121.

[24] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer, New York.

[25] Hall, P. (1993), On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation, *J. Roy. Statist. Soc., Ser. B*, 55, 291-304.

[26] Hall, P. and Wehrly, T.E. (1991). A geometrical method for removing edge effects from kernel type nonparametric regression estimators, *J. Amer. Statist. Assoc.*, vol. 86, 665-672.

[27] Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge Univ. Press.

[28] Härdle, W. and Bowman, A.W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands, *J. Amer. Statist. Assoc.*, 83, 102-110.

[29] Härdle, W. and Marron, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression, *Ann. Statist.*, 19, 778-796.

[30] Hart, J.D. (1997). *Nonparametric Smoothing and Lack-Of-Fit Tests*, Springer, New York.

[31] Hart, J.D. and Yi, S. (1998). One-sided cross-validation, *J. Amer. Statist. Assoc.*, vol. 93, no. 442, 620–631.

[32] Hong, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach, *J. Amer. Statist. Assoc.*, 94, 1201-1220.

[33] Hong, Y. and White, H. (2005). Asymptotic distribution theory for nonparametric entropy measures of serial dependence, *Econometrica*, Vol. 73, No. 3, 837-901.

[34] Horowitz, J. (1998). Bootstrap methods for median regression models *Econometrica*, vol. 66, no. 6, pp. 1327-1351.

[35] Koenker, R. (2005). *Quantile regression*, Cambridge Univ. Press.

[36] Li, Q. and Racine, J.S. (2007). *Nonparametric Econometrics*, Princeton Univ. Press, Princeton NJ.

[37] Linton, O.B., Sperlich, S. and van Keilegom, I. (2008). Estimation of a semiparametric transformation model, *Ann. Statist.*, vol.36, no. 2, pp. 686-718.

[38] Loader, C. (1999). *Local Regression and Likelihood*, Springer, New York.

[39] McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*, Chapman and Hall, London.

[40] McMurry, T. and Politis, D.N.(2008). Bootstrap confidence intervals in nonparametric regression with built-in bias correction, *Statist. Prob. Letters,* vol. 78, 2463–2469.

[41] McMurry, T. and Politis, D.N.(2010). Banded and tapered estimates of autocovariance matrices and the linear process bootstrap, *J. Time Ser. Anal.,* vol. 31, pp. 471-482.

[42] Nadaraya, E.A. (1964). On estimating regression. *Theory of Prob. Appl.*, 9, 141-142.

[43] Neumann, M. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression, *J. Nonparam. Statist.*, 9, 307-333.

[44] Olive, D.J. (2007). Prediction intervals for regression models. *Comput. Statist. and Data Anal.*, 51, pp. 3115–3122.

[45] Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*, Cambridge Univ. Press.

[46] Patel, J.K. (1989). Prediction intervals: a review, *Comm. Statist. Theory Meth.*, 18, 2393-2465.

[47] Politis, D.N. (2003). A normalizing and variance-stabilizing transformation for financial time series, in *Recent Advances and Trends in Nonparametric Statistics*, (M.G. Akritas and D.N. Politis, Eds.), Elsevier, Amsterdam, pp. 335-347.

[48] Politis, D.N. (2007a). Model-free vs. model-based volatility prediction. *J. Financial Econometrics*, vol. 5, no. 3, pp. 358-389.

[49] Politis, D.N. (2007b). Model-free prediction, in *Bulletin of the International Statistical Institute–Volume LXII*, 22 - 29 Aug. Lisbon, 2007, pp. 1391-1397.

[50] Politis, D.N. (2010). Model-free Model-fitting and Predictive Distributions, Discussion Paper, Department of Economics, Univ. of California—San Diego. Retrieved from: `http://escholarship.org/uc/item/67j6s174`

[51] Politis, D.N., Romano, J.P. and Wolf, M. (1999), *Subsampling,* Springer Verlag, New York.

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

[53] Ruppert, D. and Cline, D.H. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations, *Ann. Statist.*, 22, 185-210.

[54] Schmoyer, R.L. (1992). Asymptotically valid prediction intervals for linear models, *Technometrics*, 34, 399-408.

[55] Schucany, W.R. (2004). Kernel smoothers: an overview of curve estimators for the first graduate course in nonparametric statistics, *Statist. Sci.*, vol. 19, 663-675.

[56] Seber, G.A.F. and Lee, A.J. (2003). *Linear Regression Analysis*, Wiley, New York.

[57] Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer, New York.

[58] Shi, S.G. (1991). Local bootstrap. *Annals Inst. Statist. Math.*, 43, pp. 667-676.

[59] Stine, R.A. (1985). Bootstrap prediction intervals for regression. *J. Amer. Statist. Assoc.*, 80, 1026-1031.

[60] Tibshirani, R. (1988), Estimating transformations for regression via additivity and variance stabilization, *J. Amer. Statist. Assoc.*, 83, 394-405.

[61] Tibshirani, R. (1996), Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc., Ser. B*, vol 58, no. 1, 267-288.

[62] Wang, L., Brown, L.D., Cai, T.T. and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression, *Ann. Statist.*, 36, 646-664.

[63] Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A*, 26, 359-372.

[64] Wolfowitz, J. (1957). The minimum distance method, *Ann. Math. Statist.*, 28, 75–88.