Title
Model-free Model-fitting and Predictive Distributions

Author
Politis, Dimitris N

Publication Date
2010-03-01

Model-free model-fitting and predictive distributions

Dimitris N. Politis*
Department of Mathematics
University of California—San Diego
La Jolla, CA 92093-0112, USA

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 set-up 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, linear regression, nonparametric estimation, prediction intervals, regression, smoothing, transformations.

*This version: March 2010. Research partially supported by NSF grant DMS-07-06732. Many thanks are due to Arthur Berg, Wilson Cheung and Tim McMurry for invaluable help with R functions and computing, to Jeff Racine and Dimitrios Thomakos for helpful discussions, and to Bill Schucany for alerting the author on the undercoverage of bootstrap prediction intervals in regression some twenty years ago!

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 set-up 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 set-up 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 reported undercoverage of bootstrap prediction intervals. Furthermore, the model-free prediction principle can be viewed as a general framework for statistical inference that includes the ubiquitous Least Squares (and L_1) fitting as special cases. Finally, Appendix A provides some technical details while Appendix B brings up the notion of L_1—cross validation.

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 the 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. set-up 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 i.i.d. *sequence $\underline{\epsilon}_m^{(m)} = (\epsilon_1^{(m)}, \ldots, \epsilon_m^{(m)})'$ where each $\epsilon_i^{(m)}, i = 1, \ldots, m$ has 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 $\mathbf{X}_{n+1}$.

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. Note that 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).

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 set-up 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 $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$, under the assumed convergence in part (a). The estimator $\hat{F}_n$ 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 $\hat{g}_{n+1}$

should be plugged-in instead. 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 would 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 may be given by *resampling* methods should these be applicable in the setting at hand; see Section 2.6 for more details.

2.3 A variation of the model-free prediction principle

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 set-up 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 would be easy since in this case it would be equivalent to checking for correlation, e.g. portmanteau test, Ljung-Box, etc.

The above ideas motivate the following variation of the prediction principle that may be of particular usefulness in the case of dependent data.

Transformation into Gaussianity as a prediction stepping stone.
(a) For any integer $m \geq$ some m_o, suppose that a transformation H_m on $\mathbf{B}^m$ is found that maps the data $\underline{Y}_m = (Y_1, \ldots, Y_m)'$ into the jointly Gaussian *vector $\underline{W}_m^{(m)} = (W_1^{(m)}, \ldots, W_m^{(m)})'$ with covariance matrix V_m whose eigenvalues—viewed as sequences in m—are bounded above and below by positive constants.*
(b) Also suppose that the transformation H_m is invertible (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 $W_m^{(m)}$ alone, i.e., that

$$Y_m = \tilde{g}_m(\underline{Y}_{m-1}, \mathbf{X}_m, W_m^{(m)}) \tag{5}$$

and

$$\underline{Y}_{m-1} = \tilde{f}_m(\mathbf{X}_m;\ W_1^{(m)}, \ldots, W_{m-1}^{(m)};\ IC) \tag{6}$$

for some functions $\tilde{g}_m$ and $\tilde{f}_m$ for all $m \geq m_o$. Finally, define the vector $\underline{\epsilon}_m^{(m)} = (\epsilon_1^{(m)}, \ldots, \epsilon_m^{(m)})'$ to equal $V_m^{-1/2}\underline{W}_m^{(m)}$ where $V_m^{1/2}$ is a square root of matrix V_m. Note that $Y_m = \tilde{g}_m(\underline{Y}_{m-1}, \mathbf{X}_m, W_m^{(m)}) = \tilde{g}_m(\underline{Y}_{m-1}, \mathbf{X}_m, V_m^{1/2}\underline{\epsilon}_m^{(m)})$ which we can rename as $g_m(\underline{Y}_{m-1}, \mathbf{X}_m, \epsilon_m^{(m)})$ since the random vector $(\epsilon_1^{(m)}, \ldots, \epsilon_{m-1}^{(m)})'$ is related in a one-to-one fashion to $\underline{Y}_{m-1}$ (by induction on m).
Let F_n denote the common normal distribution of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$ that are i.i.d. by construction. Then, the L_1 and L_2–optimal model-free predictors and the predictive distribution of $g(Y_{n+1})$ given $\underline{Y}_n$ and $\mathbf{X}_{n+1}$ are given verbatim *by parts (c) and (d) of the Prediction Principle.*

In applications, the covariance matrix V_n must be estimated 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), or an appropriate shrinkage and/or regularization technique—see e.g. Bickel and Li (2006) and the references therein; then, the estimate $\hat{V}_n$ must be extrapolated to give an estimate of V_{n+1}. As before, the distribution F_{n+1} can be consistently estimated by $\hat{F}_n$, the empirical distribution of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$, or by a Gaussian distribution with unit variance and estimated mean; the former option may be more robust in practice.

Applying the Gaussian '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.

However, now that H_n is essentially a *normalizing* transformation, a collection of graphical and exploratory data analysis (EDA) tools are also available to facilitate this search. 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. In any case, these tools are often used as model-checking diagnostics in a regression context.

Remark 2.4 Note that if the normalizing transformation H_n is such that the covariance matrix V_m has diagonal elements that are (approximately) constant, then H_n deserves the name 'normalizing and variance-stabilizing' transformation[1] (NoVaS, for short). Of course, if a normalizing transformation H_n is found, then it is a matter of simple re-scaling to construct a NoVaS transformation for the data. So the Gaussian 'stepping stone' principle could equivalently have been stated insisting that the transformation H_n is also variance-stabilizing, i.e., a NoVaS transformation. Politis (2003,2007a) gives details of applying a NoVaS transformation in a setting of heteroskedastic time series, i.e., a setting like our example (2).

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

[1]This is a *data* transformation, not to be confused with the classical normalizing and variance-stabilizing transformations of *statistics* like Fisher's z tranformation for the correlation, etc.

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 set-up 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 remarkably *insensitive* to where 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 and, as a result, 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 set-up 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 practititioner 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 in the case of model-fitting, 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-fitting analogy, and use standard estimation tech-

niques such as Maximum Likelihood (ML) or Least Squares (LS) when θ is finite-dimensional, or standard nonparametric/smoothing techniques when θ is infinite-dimensional. At the end, however, the practitioner must use diagnostics and/or formal tests to ensure that the resulting values achieve the goal of the transformation, i.e., render the transformed data i.i.d. and/or Gaussian according to whether the original model-free principle or its Gaussian variation is adopted.

If the goal of the transformation is not achieved by step (A), then the strategy may be modified as follows.

(B) The parameter θ may be divided in two parts, i.e., θ=(θ_1, θ_2) where θ_1 is finite-dimensional—and, ideally, of small dimension, say 2 or 3. Firstly, θ_2 is fitted using standard methods[2] as in strategy (A). After a value for θ_2 is determined, θ_1 may be chosen as the solution to an optimization problem, i.e., as the value that renders the transformed data closest (according to some metric) to the desired goal of 'i.i.d.–ness' or Gaussianity.

Nevertheless, in certain examples the form of the desired transformation H_n is apparent; this is—fortunately—the case in the regression example analyzed in detail in Section 4.

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, i.e., possessing a smaller variance and/or inter-quartile range than ideal. The only general way to practically correct for that is via *resampling*; fortunately, the model-free principle seems ideally amenable to analysis via 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

[2]If a value for θ_1 is required in order to complete the calculation of a value for θ_2, then a preliminary value for θ_1 is obtainable from step (A).

and Tibshirani (1993), Davison and Hinkley (1997), or Shao and Tu (1995). To see how, let $\Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$ 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: $\Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n) =$

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

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:

$$g(Y_{n+1}) - \Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n) \tag{7}$$

whose distribution we can approximate by that of the bootstrap root:

$$g(Y^*_{n+1}) - \Pi(g, \hat{g}^*_{n+1}, \underline{Y}^*_n, \mathbf{X}_{n+1}, \hat{F}^*_n) \tag{8}$$

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 $\hat{g}_{n+1}(\underline{Y}^\star_n, \mathbf{X}_{n+1}, \varepsilon)$ where ε is drawn randomly from the set $(\varepsilon^{(n)}_1, ..., \varepsilon^{(n)}_n)$.

(d) Based on the pseudo-data $\underline{Y}^\star_n$, estimate the function g_{n+1} by $\hat{g}^\star_{n+1}$ respectively.

(e) Calculate a bootstrap root replicate using eq. (8).

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

$$[\Pi + q(\alpha/2),\ \Pi + q(1-\alpha/2)] \tag{9}$$

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.[3]

The above resampling algorithm is closely related to the 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 $\varepsilon^{(n)}_1, ..., \varepsilon^{(n)}_n$ 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. If $\Lambda^2(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$ is an (accurate) estimator of the variance of root (7), and $\Lambda^2(g, \hat{g}^*_{n+1}, \underline{Y}^*_n, \mathbf{X}_{n+1}, \hat{F}^*_n)$ is the corresponding estimator of the variance of the bootstrap root (8), then the predictive distribution of the studentized root

$$\frac{g(Y_{n+1}) - \Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)}{\Lambda(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)} \tag{10}$$

[3]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 $\mathbf{X}_{n+1}$.

can be approximated by that of the bootstrap root:

$$\frac{g(Y^*_{n+1}) - \Pi(g, \hat{g}^*_{n+1}, \underline{Y}^*_n, \mathbf{X}_{n+1}, \hat{F}^*_n)}{\Lambda(g, \hat{g}^*_{n+1}, \underline{Y}^*_n, \mathbf{X}_{n+1}, \hat{F}^*_n)}. \tag{11}$$

Letting $Q(\alpha)$ be the α–quantile of the empirical distribution of (11) based on B bootstrap root replicates, then a $(1-\alpha)100\%$ equal-tailed predictive interval for $g(Y_{n+1})$ of the *studentized* root type is given by

$$[\Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n) + Q(\alpha/2) \cdot \Lambda,\ \Pi(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n) + Q(1-\alpha/2) \cdot \Lambda]$$

where Λ in the above is short-hand for $\Lambda(g, \hat{g}_{n+1}, \underline{Y}_n, \mathbf{X}_{n+1}, \hat{F}_n)$. Analogously to step 5 of the Resampling Algorithm, our estimate of the predictive distribution of $g(Y_{n+1})$ would be an appropriately shifted *and* scaled version of the above empirical distribution of the B bootstrap root replicates.

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) and the references therein. Thus, in this paper we will focus on the simpler intervals of root type (9).

3 Model-based prediction in regression

3.1 Model-based nonparametric regression

We now focus on the nonparametric regression set-up of eq. (1). For simplicity, the regressor $\underline{x}_t$ will be assumed univariate and deterministic, and denoted simply as x_t. In other words, here and throughout Section 3, our data $\{(Y_t, x_t), t = 1, ..., n\}$ are assumed to have been generated by the model

$$Y_t = \mu(x_t) + \sigma(x_t)\ \varepsilon_t\ ,\quad t = 1, \ldots, n, \tag{12}$$

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

There are many approaches towards nonparametric estimation of the functions μ and σ such as 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). For simplicity of presentation, we will focus here on kernel estimators but it is important to note that the prediction procedures of this paper can equally be implemented with *any* other appropriate regression estimator, be it of parametric or nonparametric form.

The most popular form of a kernel smoother is the Nadaraya-Watson estimator (Nadaraya (1964), Watson (1964)) defined by

$$m_x = \sum_{i=1}^{n} Y_i \tilde{K}\left(\frac{x - x_i}{h}\right) \tag{13}$$

where $K(x)$ is a symmetric kernel function, and

$$\tilde{K}\left(\frac{x - x_i}{h}\right) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x-x_k}{h}\right)}. \tag{14}$$

The estimator m_x depends on the kernel K as well as on the bandwidth parameter h but this dependence will not be explicitly denoted.

Similarly, the Nadaraya-Watson estimator of $\sigma(x)$ is given by s_x defined as the (positive) square root of

$$s_x^2 = M_x - m_x^2 \quad \text{where} \quad M_x = \sum_{i=1}^{n} Y_i^2 \tilde{K}\left(\frac{x - x_i}{q}\right), \tag{15}$$

and q is another bandwidth parameter.

Selection of the bandwidth parameters h and q is usually done by (predictive) cross-validation. To elaborate, let e_t denote the *fitted* residuals, i.e.,

$$e_t = (Y_t - m_{x_t})/s_{x_t} \quad \text{for } t = 1, \ldots, n. \tag{16}$$

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

$$\tilde{e}_t = \frac{Y_t - m_{x_t}^{(t)}}{s_{x_t}^{(t)}}, \quad t = 1, \ldots, n \tag{17}$$

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

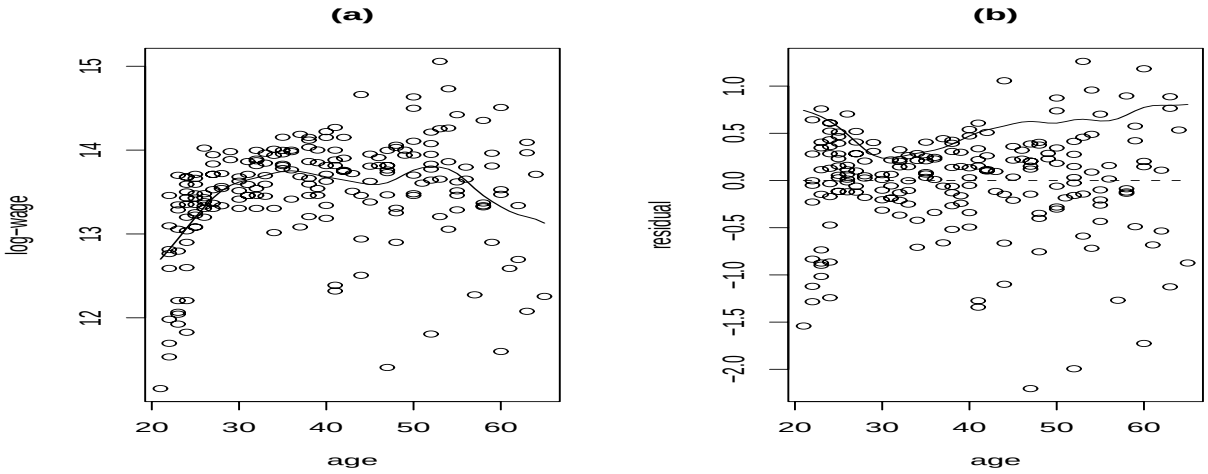

Figure 1: (a) Log-wage vs. age data with fitted kernel smoother m_x (solid line). (b) Plot of the unstudentized residuals $Y - m_x$ with superimposed estimated standard deviation s_x (solid line).

at the point x; as usual, we define $s_{xt}^{(t)} = \sqrt{M_{xt}^{(t)} - (m_{xt}^{(t)})^2}$. In other words, $\tilde{e}_t$ is the (standardized) error in trying to predict Y_t from the aforementioned delete-Y_t dataset.

Cross-validation amounts to picking the bandwidths[4] h and q 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, we instead propose using cross-validation based on an L_1 criterion; is it more robust, and is not any more computationally expensive than PRESS cross-validation. 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, to distinguish it from PRESS. In what follows in this paper, L_1—cross-validation will be used; Appendix B provides some further discussion on this choice.

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

[4]Rather than doing a two-dimensional search over h and q to minimize PRESS, the simple constraint $q = h$ will be imposed here that has the additional advantage of rendering $M_x \geq m_x^2$ as needed for a well-defined estimator s_x^2 in eq. (15). Note, however, 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.

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 for smoothing. 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). Nevertheless, we will not elaborate further here since our purpose is to develop general prediction procedures that 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 $Y_{\rm f}$ associated with a potential design point $x_{\rm f}$. Recall that the L_2–optimal (point) predictor of $Y_{\rm f}$ is the expected value of the response $Y_{\rm f}$ associated with design point $x_{\rm f}$ which will be denoted $E(Y_{\rm f}|x_{\rm f})$; under model (12), we have that $E(Y_{\rm f}|x_{\rm f}) = \mu(x_{\rm f})$. 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 $Y_{\rm f}$ associated with design point $x_{\rm f}$; under model (12), this is given by $\mu(x_{\rm f}) + \sigma(x_{\rm f}) \cdot median(F)$. If the error distribution F is symmetric around zero, 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, $\mu(x_{\rm f})$ and $\sigma(x_{\rm f})$ are estimated by $m_{x_{\rm f}}$ and $s_{x_{\rm f}}$ of eq. (13) and (15). 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. (16). Hence, the practical L_2— and L_1—optimal *model-based* predictors of $Y_{\rm f}$ are given respectively by $\widehat{Y}_{\rm f} = m_{x_{\rm f}}$ and $\widetilde{Y}_{(x)} = m_{x_{\rm f}} + s_{x_{\rm f}} \cdot median(\hat{F}_e)$.

Suppose, however, that our objective is predicting the future value $g(Y_{\rm f})$ associated with design point $x_{\rm f}$ where $g(\cdot)$ is a function of interest; this possibility is of particular importance 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 (12). For example, the wages in dataset `cps71` have been logarithmically transformed before model (12) 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

$$n^{-1}\sum_{i=1}^{n} g\left(m_{x_{\rm f}} + \sigma_{x_{\rm f}} e_i\right).$$

Note that the naive predictor $g(m_{x_{\rm f}})$ can be grossly suboptimal when g is appreciably nonlinear. Similarly, the model-based L_1–optimal (point) predictor of $g(Y_{\rm f})$ can be approximated by the sample median of the set $\{g\left(m_{x_{\rm f}} + \sigma_{x_{\rm f}} e_i\right),\ i = 1, ..., n\}$.

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. (12) for his/her statistical analysis. However, even while fitting—and working with—model (12), 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.

Recall that throughout Section 3 we assume that—typically unbeknownst to the statistician—model (12) is true. Hence, the model-free practitioner should find (via the usual diagnostics) that to a good approximation the fitted residuals $e_t = (Y_t - m_{x_t})/s_{x_t}$ are close to being i.i.d. However, the model-free practitioner does not see this as model confirmation but as a good starting point for the model-free principle as suggested by Remark 2.3.

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 local polynomial fitting and other popular methods. Motivated by the studentizing transformation in Politis (2003,2007a), we can use the linearity of m_x and consider a more general centering/studentization that may provide a better transformation for the model-free principle. Such a transformation is given by:

$$W_t = \frac{Y_t - \tilde{m}_{x_t}}{\tilde{s}_{x_t}},\ t = 1, \ldots, n. \tag{18}$$

where

$$\tilde{m}_{x_t} = cY_t + (1-c)m_{x_t}^{(t)},\ \tilde{M}_{x_t} = cY_t^2 + (1-c)M_{x_t}^{(t)} \ \text{ and } \ \tilde{s}_{x_t}^2 = \tilde{M}_{x_t} - \tilde{m}_{x_t}^2\ . \quad (19)$$

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 will not be explicitly denoted. Details on the choice of parameter c will be given later.

Eq. (18) is a more general—and thus more flexible—reduction to residuals since it includes the fitted residuals (16) as a special case. To see this, note that (13) 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. (18) reduces to eq. (16). The generality of eq. (18) is further shown by considering different options for c. For example, consider the extreme case of $c = 0$; in this case, W_t is tantamount to a *predictive* residual, i.e., $W_t = \tilde{e}_t$ defined in eq. (17).

Thus, eq. (18) 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 the design parameters (c and the bandwidth), $W_1, \ldots, W_n$ 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 (12) 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 effect, note that eq. (18) can still be written for the unobserved

[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.

$Y_{\rm f}$, i.e., the yet unobserved $Y_{\rm f}$ is related to the yet unobserved $W_{\rm f}$ by

$$W_{\rm f} = \frac{Y_{\rm f} - \tilde{m}^{\rm f}_{x_{\rm f}}}{\tilde{s}^{\rm f}_{x_{\rm f}}} \tag{20}$$

where $\tilde{m}^{\rm f}$ and $\tilde{s}^{\rm f}$ are the estimators from eq. (13) and (15) 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. (19) we have:

$$\tilde{m}^{\rm f}_{x_{\rm f}} = cY_{\rm f} + (1-c)m_{x_{\rm f}},\ \tilde{M}^{\rm f}_{x_{\rm f}} = cY^2_{\rm f} + (1-c)M_{x_{\rm f}} \ \text{ and } \ \tilde{s}^{\rm f}_{x_{\rm f}} = \sqrt{\tilde{M}^{\rm f}_{x_{\rm f}} - (\tilde{m}^{\rm f}_{x_t})^2} \tag{21}$$

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. (20) for $Y_{\rm f}$ is the key to model-free prediction as it would yield an equation like (3). As verified in the Appendix, the solution of eq. (20) is given by

$$Y_{\rm f} = m_{x_{\rm f}} + s_{x_{\rm f}} \frac{W_{\rm f}}{\sqrt{1 - c - cW^2_{\rm f}}}. \tag{22}$$

Eq. (22) 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. As a matter of fact, the transformation of Y to W is a kurtosis-reducing transformation. As can easily be verified, the (sample) kurtosis of $W_1, ..., W_n$ is a continuous function of c that tends to zero when $c \to 1$. So, by the intermediate value theorem, there is an appropriate choice of $c \in [0, 1)$ that makes the (sample) kurtosis of $W_1, ..., W_n$ match any desired value in $(0, \tilde{k})$ where $\tilde{k}$ is an estimate of the kurtosis of the Y's. In particular, if the Y data are heavy-tailed with approximately symmetric distribution, then an appropriate choice of c would make the kurtosis of $W_1, ..., W_n$ equal to the Gaussian value of 3; in that case, the transformation of Y to W would be a *normalizing* transformation—at least as regards the first four moments.

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. (22) depend on the value of c. Surprisingly (and thankfully), the answer

is *not at all* ! To see this, note that after some algebra:

$$\frac{W_t}{\sqrt{1-c-cW_t^2}} \equiv \tilde{e}_t \quad \text{for any} \quad c \in [0,1), \quad \text{and for all} \quad t = 1, \ldots, n, \tag{23}$$

where the $\tilde{e}_t$s are the *predictive* residuals defined in eq. (17). In other words, the prediction equation (22) does *not* depend on the value of c, and can be simplified to:

$$Y_{\mathrm{f}} = m_{x_{\mathrm{f}}} + s_{x_{\mathrm{f}}}\tilde{e}_{\mathrm{f}}. \tag{24}$$

Eq. (24) will form the basis for our application of the model-free prediction principle under model (12). Since the model-free philosophy is implemented in a set-up where model (12) 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. (24) 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. (24). Note, however, that the only difference between eq. (24) 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}	$m_{x_{\mathrm{f}}}$	$m_{x_{\mathrm{f}}} + s_{x_{\mathrm{f}}} \cdot \mathrm{mean}(\tilde{e}_i)$
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 (12) is true.

3.5 Model-free/model-based prediction intervals

Note that the model-based L_2—optimal predictor of $Y_{\rm f}$ 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 is to center the residuals e_i to their mean, and use the centered residuals for prediction; the necessity of centering of the residuals was first pointed out by Freedman (1981), and will also be used in the Resampling Algorithm in what follows.

Of course, 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

$$\tilde{e}_t = e_t/(1-\delta_{x_t}) \;\; \text{where } \delta_{x_t} = K(0)/\sum_{k=1}^{n} K\left(\frac{x_t - x_k}{h}\right). \tag{25}$$

Since $h \to 0$ as $n \to \infty$, it follows that $\delta_{x_t} \to 0$, i.e., the model-free and model-based predictors are *asymptotically* equivalent in the regression example. Nevertheless, since $\delta_{x_t} > 0$ for any finite n, $\tilde{e}_t$ will always be *larger* in absolute value (i.e., inflated) as compared to e_t, and this may make a difference in practice.

Eq. (25) suggests that the main difference between the fitted and predictive residuals is their scale; their center should be about the same (and close to zero). Therefore, the model-based and MF/MB *point* predictors of $Y_{\rm f}$ are almost indistinguishable; this is, of course, reassuring since, when model (12) 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 $Y_{\rm f}$ in which case the correct scaling of residuals is of paramount importance; see also the discussion in Section 3.7. With regards to the construction of an accurate predictive distribution of $Y_{\rm f}$, 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 $\sigma(x)$.

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

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

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 $\Pi = n^{-1}\sum_{i=1}^n g\left(m_{x_{\rm f}} + \sigma_{x_{\rm f}} r_i\right)$

   (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 $Y_{\rm f}^\star = m_{x_{\rm f}} + s_{x_{\rm f}} r$ 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 $g(Y_{\rm f}^\star) - \Pi(g, m_x^*, s_x^*, \underline{Y}_n^*, \mathbf{X}_{n+1}, \hat{F}_n^*)$.

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 whose α—quantile is denoted $q(\alpha)$.

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

$$[\Pi + q(\alpha/2), \Pi + q(1-\alpha/2)]. \tag{26}$$

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.

Remark 3.1 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 3.2 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 4.7 for more discussion.

Remark 3.3 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 (12). In that case, m_{x_f} is also normal, and independent of the 'future' error ε_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 $Y_\mathrm{f} - m_{x_\mathrm{f}}$ implies an approximate $(1-\alpha)100\%$ equal-tailed, predictive interval for Y_f given by:

$$[m_{x_\mathrm{f}} + V_{x_\mathrm{f}} \cdot z(\alpha/2),\ m_{x_\mathrm{f}} + V_{x_\mathrm{f}} \cdot z(1-\alpha/2)] \tag{27}$$

where $V_{x_\mathrm{f}}^2 = s_{x_\mathrm{f}}^2 \left(1 + \sum_{i=1}^n \tilde{K}^2(\frac{x_\mathrm{f}-x_i}{h})\right)$ with $\tilde{K}$ defined in eq. (14), 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

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

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

Interval (27) 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 (27) is quite limited.

3.6 Fitting parametric regression via the MF/MB paradigm

In this subsection, we show how the model-free principle can be applied to fit a parametric model when such a model is assumed. To fix ideas, consider the simple straight-line regression set-up where $Y_i = \beta_0 + \beta_1 x_i + Z_i$ with $Z_i \sim$ i.i.d. $(0,\sigma^2)$ for $i = 1, \ldots, n$. The essence of the above model—as far as model-free prediction is concerned—is that $\eta_i \equiv Y_i - \beta_1 x_i$ are i.i.d. albeit with (possibly non-zero) mean β_0.

Thus, a candidate transformation to 'i.i.d.–ness' by $r_i = Y_i - \hat{\beta}_1 x_i$ where $\hat{\beta}_1$ is a candidate value. The model-free principle mandates choosing $\hat{\beta}_1$ with the objective of having the r_is become as close to i.i.d. as possible. However, under the linear regression model, the r_is would be i.i.d. if only their first moment was properly adjusted. So, in this case, the model-free principle suggests choosing choosing $\hat{\beta}_1$ in such

a way to make $r_1, \ldots, r_n$ have (approximately) the *same* first moment. Noting that the latter is well approximated by the empirical value $\hat{r} = n^{-1}\sum_{i=1}^{n} r_i$, we can use a subsampling construction to make this happen.

To start with, assume that the design points $x_1, ..., x_n$ are fixed (nonrandom), and sorted in ascending order. Following the construction in Politis, Romano and Wolf (1999, Ch. 9.2) using partially overlapping blocks of size b, compute the block means

$$\bar{r}_{k,b} = b^{-1} \sum_{t=L(k-1)+1}^{L(k-1)+b} r_t \quad \text{for} \quad k = 1, ..., q \tag{28}$$

where $q = [L^{-1}(n-b)] + 1$ and $[\cdot]$ is the integer part. Here L indicates the degree of overlap of the blocks; with $L = b$ we have non-overlapping blocks, whereas with $L = 1$ the overlap is the maximum possible—the latter is recommended if it is computationally feasible.

Note that $\bar{r}_{k,b}$ is an estimate of the first moment of the r_is found in the kth block. Thus, the requirement that all $r_1, \ldots, r_n$ have first moment (approximately) equal to $\hat{r}$ can be written formally as follows:

$$\text{Choose } \hat{\beta}_1 \text{ that minimizes } LS(b) = \sum_{k=1}^{q} (\bar{r}_{k,b} - \hat{r})^2 \text{ or } L1(b) = \sum_{k=1}^{q} |\bar{r}_{k,b} - \hat{r}| \tag{29}$$

according to whether an L_2 or L_1 criterion of closeness is preferred.

In contrast to the use of subsampling for variance or distribution estimation, it is not necessary here that b is large. Even the value $b = 1$ is plausible in which case we have:

$$\frac{d}{d\hat{\beta}_1} LS(1) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{where } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

In other words, the model-free procedure (29) with L_2 criterion and $b = 1$ is reassuringly *identical* to the usual Least Squares estimator! Now the r_is serve as proxies for the unobservable η_is which have expected value β_0 under the model; hence, β_0 is naturally estimated by the sample mean of the r_is, i.e.,

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 x_i) = \bar{Y} - \hat{\beta}_1 \bar{x}$$

which is again the Least Squares estimator.

Note that minimizing $LS(b)$ with $b > 1$ gives a more robust way of doing Least Squares in which the effect of outliers is diminished by the local averaging of b neighboring values; we do not elaborate further here due to lack of space. Similarly to the above, minimizing $L1(1)$ is equivalent to L_1 regression, whereas minimizing $L1(b)$ with $b > 1$ is an even more robust procedure.

Remark 3.4 In all the above, the block mean $\bar{r}_{k,b}$ of eq. (28) could be replaced by the (sample) *median* of the block $\{r_t$, for $t = L(k-1)+1, \ldots, L(k-1)+b\}$. The resulting minimization of $LS(1)$ or $L1(1)$ is still equivalent to Least Squares or L_1 regression respectively while the minimization of $LS(b)$ or $L1(b)$ with $b > 1$ gives some different variation of robust regression.

In concluding, we now outline the general case of fitting a parametric regression via the model-free principle. Consider the model $Y_i = f_\theta(x_i) + Z_i$ with $Z_t \sim$ i.i.d. (β_0, σ^2) for $i = 1, \ldots, n$; here, f_θ belongs to a parametric family indexed by the finite-dimensional parameter θ. We again assume that the design points $x_1, ..., x_n$ are fixed (nonrandom), and sorted in ascending order. Let $\hat{\theta}$ be a candidate value, and define $r_i = Y_i - f_{\hat{\theta}}(x_i)$ with $\hat{r} = n^{-1}\sum_{i=1}^n r_i$ as before. Letting $\bar{r}_{k,b}$ denote the sample mean (or median) of the block $\{r_t,\ t = L(k-1)+1, \ldots, L(k-1)+b\}$, the MF/MB fitting procedure amounts to

$$\text{choosing } \hat{\theta} \text{ that minimizes } LS(b) = \sum_{k=1}^{q} (\bar{r}_{k,b} - \hat{r})^2 \text{ or } L1(b) = \sum_{k=1}^{q} |\bar{r}_{k,b} - \hat{r}| \qquad (30)$$

according to whether an L_2 or L_1 criterion is preferred. Finally, estimate β_0 and σ^2 by the sample mean and sample variance of the r_is respectively.

Finally, note that in all the above—and in eq.

3.7 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 literature 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:

$$Y_i = \underline{x}_i'\underline{\beta} + Z_i, \text{ for } i = 1, \ldots, n, \tag{31}$$

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 having the $p \times 1$ vector $\underline{x}_i'$ as its ith row.

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. (24), 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_t 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. (25), 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

$$\tilde{z}_i = \frac{z_i}{1 - h_i}, \text{ for } i = 1, \ldots, n, \tag{32}$$

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 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 (accurate) estimator of $\sigma^2(\underline{x}_i)$ is available via parametric or nonparametric methods.

3.8 Simulation: better prediction intervals in linear regression

We now conduct a small simulation in the linear regression set-up of subsection 3.7 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 $x_f = 1$. The simulation focused on constructing 90% prediction intervals, and was based on 900 repetitions of each experiment. Figure 2 shows two typical scatterplots with superimposed Least Squares (LS) line; both LS regression and L_1 regression were considered for estimating β_0 and β_1.

Table 3.2 reports the empirical coverage levels (COV), and (average) lower and upper limits of the different prediction intervals in the linear regression case. The standard error of the COV 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 $x_f = 1$ is zero; so the different prediction intervals are expected to be centered around zero. Indeed, all (average) intervals of Table 3.2 are approximately symmetric around zero.

[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.

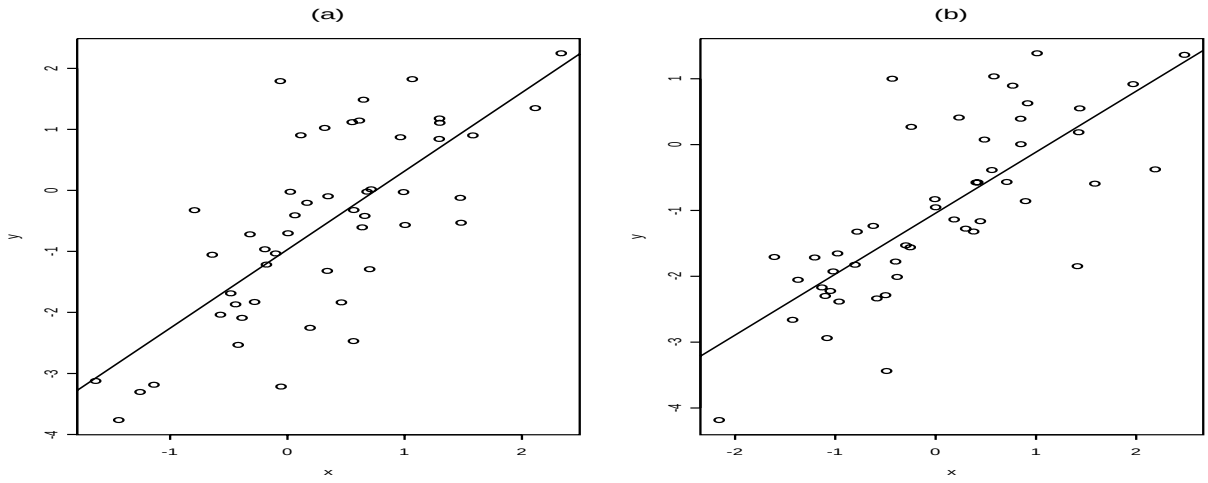

Figure 2: Typical linear regression scatterplots with superimposed Least Squares lines; (a) Normal data; (b) Laplace data.

Linear regression is, of course, a model-based set-up; 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 $m_{x_\mathrm{f}} \pm t_{n-2}(\alpha/2)S\sqrt{1+h_\mathrm{f}}$; and Olive's (2007) 'semi-parametric' interval:

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

In the above, m_{x_f} is the usual point predictor given by $\hat{\beta}_0 + \hat{\beta}_1 x_\mathrm{f}$, $h_\mathrm{f} = \underline{x}_\mathrm{f}'(X'X)^{-1}\underline{x}_\mathrm{f}$ is the 'leverage' at point x_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., COV(MB) < COV(MB Stine) < COV(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 $x_f = 1$	COV	INTERVAL	(st.err.)	COV	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 (COV), 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 regresion

4.1 Constructing the transformation

We now revisit the nonparametric regression set-up of Section 3 but in a situation where a model such as eq. (12) can not be considered to hold true (not even approximately). As an example of model (12) not being valid, consider the set-up 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 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 3.

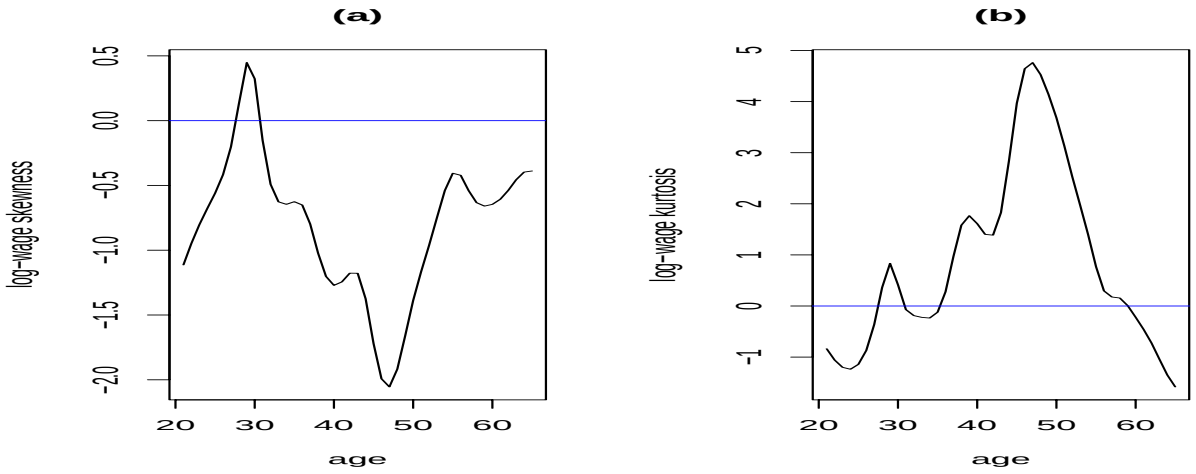

Figure 3: (a) Skewness of log-wage vs. age. (b) Kurtosis of log-wage vs. age. [Kernel-based estimates from dataset `cps71`.]

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), or McCulloch (2000). 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

$$N_{x,h}^{-1} \sum_{t:|x_t - x| < h/2} \mathbf{1}\{Y_t \leq y\} \tag{33}$$

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

$$\hat{D}_x(y) = \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\} \tilde{K}\left(\frac{x - x_i}{h}\right) \tag{34}$$

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. (33) is just $\hat{D}_x(y)$ with K chosen as the rectangular kernel, i.e., $K(x) = \mathbf{1}\{|x| \leq h/2\}$; in general, we can use any non-negative, integrable kernel $K(\cdot)$ in (34).

Estimator $\hat{D}_x(y)$ enjoys many good properties including asymptotic consistency under some conditions; see e.g. Theorem 6.1 of Li and Racine (2007). Nevertheless, it is discontinuous as a function of y, and therefore unacceptable for our purposes. To come up with an estimator that is continuous (and strictly increasing) in y we propose the following construction.[8]

For x fixed, $\hat{D}_x(y)$ is a step function with (possible) jumps at the data points $Y_1, \ldots, Y_n$. However, some data points receive zero weight in (34) being far away from the x location in question. As before, suppose there are $N_{x,h}$ data points receiving positive weight in (34), i.e., $N_{x,h} = \#\ \{t : K(\frac{x-x_t}{h}) > 0\}$. Assuming $N_{x,h} > 1$, we order these $N_{x,h}$ data points in increasing order and denote them by $Y^{(x)}_{[1]} < Y^{(x)}_{[2]} < \cdots < Y^{(x)}_{[N_{x,h}]}$. Now let $A_1, \ldots, A_{N_{x,h}-1}$ denote the midpoints of the step 'sizes' of the step function $\hat{D}_x(y)$, i.e., let $A_i = (Y^{(x)}_{[i]} + Y^{(x)}_{[i+1]})/2$ for $i = 1, \ldots, N_{x,h}-1$. To complete the construction we have to define A_0 and $A_{N_{x,h}}$; a conservative choice is $A_0 = Y^{(x)}_{[1]}$ and $A_{N_{x,h}} = Y^{(x)}_{[N_{x,h}]}$ but in what follows the symmetric assignment $A_0 = 2Y^{(x)}_{[1]} - A_1$ and $A_{N_{x,h}} = 2Y^{(x)}_{[N_{x,h}]} - A_{N_{x,h}-1}$ will be used. Finally, linear interpolation between the points $A_0, A_1, \ldots, A_{N_{x,h}}$ gives our continuous *and* strictly increasing (in y) estimator that will be denoted by $\tilde{D}_x(y)$; Figure 4 (a) exemplifies this construction.

Remark 4.1 For $\tilde{D}_x$ 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.

[8] A smooth (differentiable in y) version of $\hat{D}_x(y)$ can be concocted in the usual way by integrating a kernel estimator of the underlying density; see Section 6 of Li and Racine (2007) for details. However, the resulting estimator of $D_x(y)$ will not be almost surely strictly increasing in y unless a kernel K of infinite support is employed. In addition, we have little use for (nor assume) differentiability of $D_x(y)$ in y here.

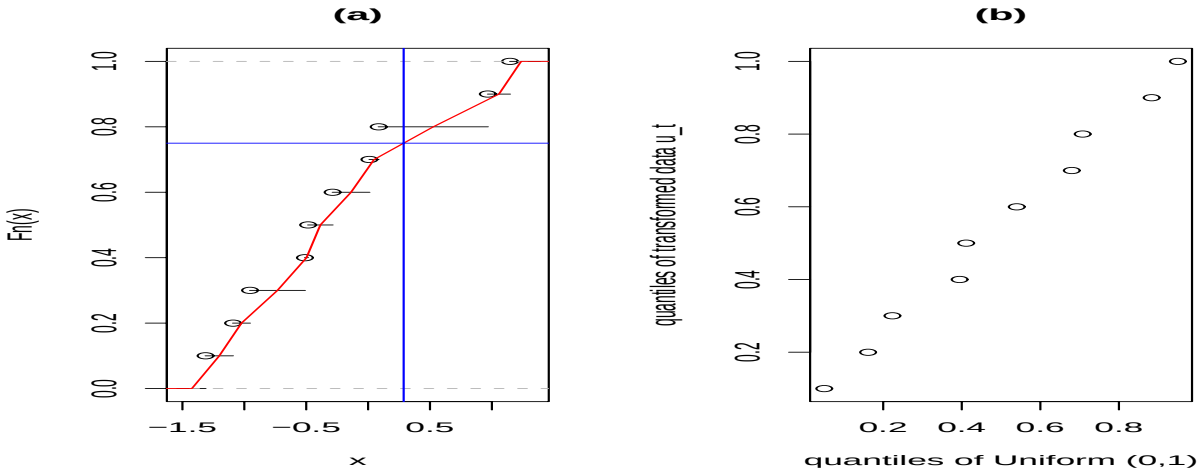

Figure 4: (a) Empirical distribution of a test sample consisting of five $N(0,1)$ and five $N(1/2,1)$ independent r.v.'s with the piecewise linear estimator $\tilde{D}(\cdot)$ superimposed; the vertical/horizontal lines indicates the inversion process, i.e., finding $\tilde{D}^{-1}(0.75)$. (b) Q-Q plot of the transformed variables u_i vs. the quantiles of Uniform (0,1).

Remark 4.2 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 variable bandwidth might be advisable in connection with the construction of $\hat{D}_x$ and $\tilde{D}_x$. Alternatively, a *k–nearest neighbor* technique may be used; in this case, the form of $\hat{D}_x$ and $\tilde{D}_x$ remains the same but the bandwidth h is taken as the (Euclidean) distance of between x and its kth nearest neighbor among $x_1, ..., x_n$. The result is a 'local' bandwidth, i.e., a bandwidth that depends on x; see e.g. Li and Racine (2007, Ch. 14). In addition, a local linear (or polynomial) smoother of the variables $\mathbf{1}\{Y_t \leq y\}$ could be used in place of the local constant estimator (34), and may be preferable because of better handling of edge effects as well as non-equally spaced x–points; details can be found in Li and Racine (2007, Ch. 6) but the essence of our discussion here remains unchanged.

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 $\tilde{D}_x(\cdot)$ as its proxy. Therefore, our proposed transformation amounts to defining

$$u_i = \tilde{D}_{x_i}(Y_i) \quad \text{for} \quad i = 1, \ldots, n; \tag{35}$$

by the consistency of $\tilde{D}_x(\cdot)$, we can now claim that $u_1, \ldots, u_n$ are approximately i.i.d. Uniform(0,1). Figure 4 (b) shows that this claim is plausible even with a sample size of just ten independent r.v.'s that are only *approximately* identically distributed as in the nonparametric regression case.

Remark 4.3 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.

Remark 4.4 If there is some suspicion of non-independence of the Y_ts, then the Gaussian 'stepping stone' may be useful. To elaborate, one would let $Z_t = \Phi^{-1}(u_t)$ for $t = 1, \ldots, n$ where Φ is the distribution of a standard normal. Then, one would examine (an estimate of) the covariance matrix of $\underline{Z}_n = (Z_1, ..., Z_n)$ to diagnose a possible non-independence.

The probability integral trasform has been used in the past as an intermediate step towards building better density estimators; see e.g. Ruppert and Cline (1994). 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. (35), 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 $\tilde{D}_x^{-1}$ which is well-defined since $\tilde{D}_x(\cdot)$ is strictly increasing by construction. Note that, for any $i = 1, ..., n$, $\tilde{D}_{x_\mathrm{f}}^{-1}(u_i)$ is a *bona fide* potential response Y_f associated with predictor x_f since $\tilde{D}_{x_\mathrm{f}}^{-1}(u_i)$ has (approximately) the same distribution as Y_f. These n valid potential responses given by $\{\tilde{D}_{x_\mathrm{f}}^{-1}(u_i)$ for $i = 1, ..., n\}$ can be

gathered together to give us an approximate empirical distribution for $Y_{\rm f}$ 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 $g(Y_{\rm f})$ will be the expected value of $g(Y_{\rm f})$ that is approximated by*

$$n^{-1}\sum_{i=1}^{n} g\left(\tilde{D}_{x_{\rm f}}^{-1}(u_i)\right). \tag{36}$$

Similarly, the L_1—optimal predictor of $g(Y_{\rm f})$ will be approximated by the sample median of the set $\{g\left(\tilde{D}_{x_{\rm f}}^{-1}(u_i)\right),\ i=1,...,n\}$. The model-free predictors[9] are summarized in Table 4.1 that can be compared to Table 3.1 of the previous section.

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

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

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 case and that $g(x) = x$. Calculating the value of the optimal predictor of eq. (36) for many different $x_{\rm f}$ values—say taken on a grid over the range of the original predictors $x_1, ..., x_n$—, 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 (12),

[9]For $\tilde{D}_{x_{\rm f}}^{-1}$ to be an accurate estimator of $D_{x_{\rm f}}^{-1}$, the value $x_{\rm f}$ 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 outside the range of the original predictors $x_1, ..., x_n$, i.e., extrapolation, is *not* feasible in the model-free paradigm.

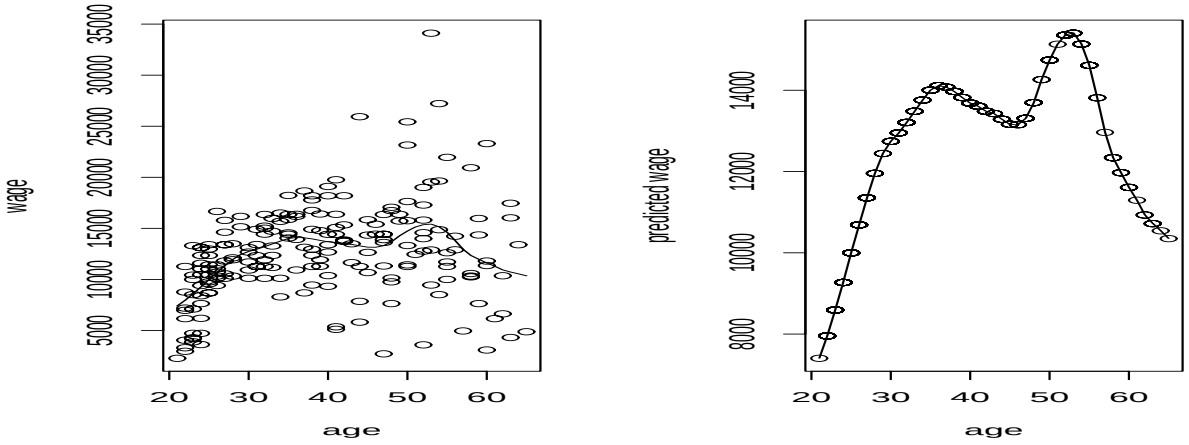

Figure 5: (a) Wage vs. age scatterplot. (b) Circles indicate the salary predictor from eq. (36) calculated from log-wage data with $g(x)$ exponential. For both figures, the superimposed solid line represents the MF^2 salary predictor calculated from the *raw* data (without the log-transformation).

therefore also qualifying to be called a model-free (point) predictor. Predictor (36) can then be seen as an alternative method to estimate $E(Y_{\rm f}|x_{\rm f})$; although it is not identical to $m_{x_{\rm f}}$, it tends to give results very close to it in practice—as one would hope since both methods are consistent for $E(Y_{\rm f}|x_{\rm f})$ under standard assumptions. For example, Figure 1 (a) looks exactly the same when the curve obtained from predictor (36) is used in place of the kernel smoother m_x since the relative difference between the two smooth curves is less than 0.1% for the log-wage vs. age dataset.

The real advantages of the model-free philosophy, however, are twofold: (a) it gives us the opportunity 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 $m_{x_{\rm f}}$ without resort to a model like (12); and (b) it is a totally *automatic* method that does not require any preliminary preprocessing and/or data transformations—see Remark 4.5 below.

Remark 4.5 The model-free prediction technique based on transformation (35) relieves the practitioner from the need to find an optimal transformation for additivity and variance stabilization such as the Box/Cox power family, ACE and/or AVAS; see Linton et al. (1997) and the references therein. Figure 5 (a) is the analog of

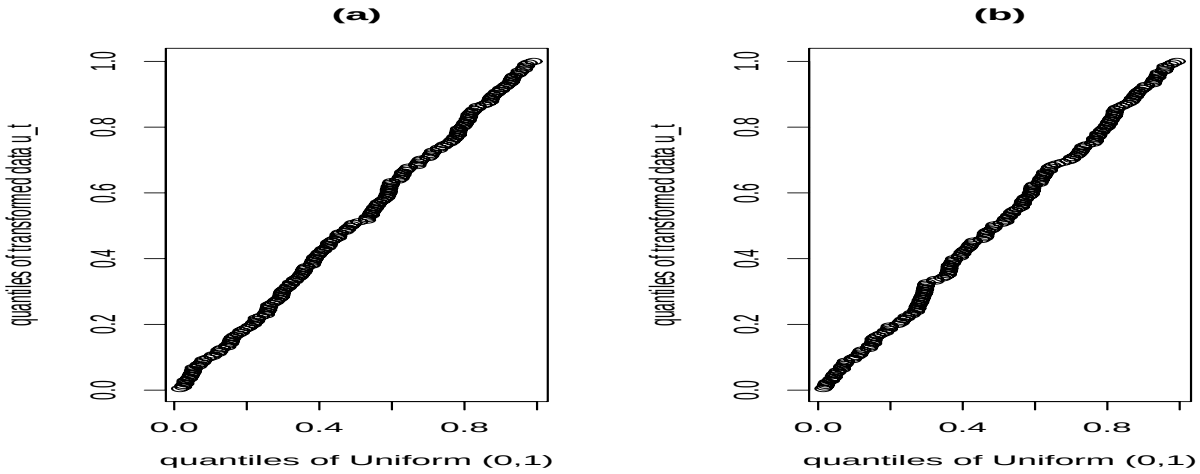

Figure 6: Q-Q plots of the transformed variables u_i vs. the quantiles of Uniform (0,1) for the Canadian wage/age dataset. (a) The u_i's are obtained from the log-wage vs. age dataset of Figure 1 using bandwidth 5.5; (b) The u_i's are obtained from the raw (untransformed) dataset of Figure 5 using bandwidth 7.3.

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 (35) on the raw data; as Figure 5 (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 (36). Figure 6 (a) shows the Q-Q plot of the transformed variables u_i based on the logarithmically transformed data whereas Figure 6 (b) is its analogue based on the raw data; in both cases, the uniformity seems to be largely achieved. Note, however, that the cross-validated optimal bandwidth choice is different in these two cases; the next subsection elaborates upon this phenomenon.

4.3 Cross-validation for model-free prediction

As seen in the last two subsections, estimating the conditional distribution $D_x(\cdot)$ by $\tilde{D}_x(\cdot)$ 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. (36), 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 $\tilde{D}_x^{(t)}(\cdot)$ computed from the delete-Y_t dataset, and having only $n-1$ values of u_i in connection with eq. (35) and (36). Finally, define the MF^2 predictive residuals

$$\tilde{e}_t = g(Y_t) - \Pi_t^{(t)} \quad \text{for} \quad t = 1, \ldots, n. \tag{37}$$

Choosing the best bandwidth h to use in our model-free predictor (36) 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 $\tilde{D}_x$ 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. (37) can be quite computationally expensive. In view of the discussion in the previous subsection argueing that the L_2—optimal predictor of eq. (36) 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; see Appendix B for more details.

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 $\tilde{D}_x$; 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 $\Pi = n^{-1} \sum_{i=1}^n g\left(\tilde{D}_{x_{\rm f}}^{-1}(u_i)\right)$ or $\Pi = \text{median}\,\{g\left(\tilde{D}_{x_{\rm f}}^{-1}(u_i)\right)\}$. 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 $\tilde{D}_x(\cdot)$, and

use eq. (35) 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 $\tilde{D}_x^{-1}$ to create pseudo-data in the Y domain, i.e., let $\underline{Y}_n^\star = (Y_1^\star, ..., Y_n^\star)$ where $Y_t^* = \tilde{D}_{x_t}^{-1}(u_t^*)$.

(c) Generate a bootstrap pseudo-response $Y_{\rm f}^*$ by letting $Y_{\rm f}^* = \tilde{D}_{x_{\rm f}}^{-1}(u)$ 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 $\tilde{D}_x^*(\cdot)$.

(e) Calculate a replicate of the bootstrap root $g(Y_{\rm f}^*) - \Pi^*$ where $\Pi^* = n^{-1}\sum_{i=1}^n g\left(\tilde{D}_{x_{\rm f}}^{*^{-1}}(u_i^*)\right)$ or $\Pi^* = \text{median}\,\{g\left(\tilde{D}_{x_{\rm f}}^{*^{-1}}(u_i^*)\right)\}$ 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 whose α—quantile is denoted $q(\alpha)$.

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

$$[\Pi + q(\alpha/2), \Pi + q(1-\alpha/2)] \tag{38}$$

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 4.6 To further build on Remark 4.5, note that the above model-free prediction interval is *invariant* with respect to the choice of function $g(\cdot)$ in a way analogous to the transformation invariance property of bootstrap confidence intervals of percentile type. To elaborate, if either point or interval prediction of $g(Y_{\rm f})$ is desired, then the model-free techniques can be immediately applied to the $\{(g(Y_t), \underline{x}_t),\ t = 1, \ldots, n\}$ dataset without worrying about how the scatterplot of $g(Y)$ vs. x looks. For example, if the objective is prediction of `wage` for a certain age group as in our `cps71` dataset, the regression would simply be `wage` vs. `age` and the need for the

log-transformation is obliterated. Consequently, the model-free prediction scheme in regression is a totally *automatic* technique.

Remark 4.7 Smoothing techniques are often plagued by edge effects. As previously mentioned, this is especially true for kernel smoothers; local linear and local polynomial estimators are much preferable in that respect. Hopefully, the future point of interest x_f will not be a boundary point in which case it may be advisable to *omit* the u_is that are obtained from x_is that are close to the boundary; for example, both Figures 1(a) and 5(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.[10] Note that a full-size dataset $(Y_1^*, ..., Y_n^*)$ can (and should) be re-created in Step 1(b) of the Resampling Algorithm even though we are using just the u_is that are away from either boundary (say there are m of these); to do this, a bootstrap with *larger* resample size is employed, i.e., based on a u—dataset of size m, a bootstrap resample $(u_1^*, ..., u_n^*)$ of size n is generated. Based on the full size pseudo-sample $(Y_1^*, ..., Y_n^*)$, we compute the bootstrap estimator $\tilde{D}_x^*(\cdot)$; however, only the Y^*s that are away from the boundaries (m in number) will be used in the construction of Π^* in Step 1(e) of the Resampling Algorithm.

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 the MB method does— typically results in variance underestimation and undercoverage of prediction intervals.

Since MF^2 predictive residuals are computable from eq. (37), 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. (37) 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.

[10]The same recommendation also applies to the MB and MF/MB of Section 3: for either point or interval predictors, only include the e_is and/or $\tilde{e}_i$s obtained from x_is that are away from either boundary by more than half a bandwidth.

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 a genuinely "outside" point. We can still achieve this effect within the MF^2 paradigm using an analogous trick; to see how, let $\tilde{D}_{x_t}^{(t)}$ denote the estimator $\tilde{D}_{x_t}$ 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

$$u_t^{(t)} = \tilde{D}_{x_t}^{(t)}(Y_t) \quad \text{for} \;\; t = 1, \ldots, n; \tag{39}$$

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 (12) holds true, the $u_t^{(t)}$s are approximately i.i.d. in general under the weak assumptions of smoothness and continuity 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 (12) 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 Problems and diagnostics

The model-free prediction scheme in regression has been developed under minimal assumptions including continuity of $D_x(y)$ in both x and y, and availability of enough data so that 'local' estimation can take place. With regards to the latter, traditional conditions for asymptotic validity would include the usual requirement that $h \to 0$ as $n \to \infty$ but also ensuring $N_{x,h} \to \infty$ for all x over an interval of interest; see

Remark 4.1. For good finite-sample results, however, we would like $\tilde{D}_x(\cdot)$ to remain largely unchanged over an interval of length $2h$ where h is the chosen bandwidth in the practical application.

With regards to the requirement of continuity of $D_x(y)$ in y, consider the extreme example where $Y = \beta_0 + \beta_1 x$ exactly (no random error), and assume an equi-spaced design on the x axis. Here, Y (given x) has a distribution that is degenerate having a point mass of unity at $\beta_0 + \beta_1 x$; hence, the continuity assumption for $D_x(y)$ breaks down and complications ensue.

To elaborate, let x be a point not on the boundary; since h must be big enough so that $N_{x,h}$ is appreciable, it follows that our $\hat{D}_x(\cdot)$ will be a discrete uniform distribution with center at $\beta_0 + \beta_1 x$ and range dictated by the parameter h. By the linearization, $\tilde{D}_x(\cdot)$ will be a continuous uniform distribution with same center and range. Therefore,

$$u_i = \tilde{D}_{x_i}(Y_i) = \tilde{D}_{x_i}(\beta_0 + \beta_1 x_i) = 1/2 \tag{40}$$

for any i such that x_i is not on the boundary, since $\beta_0 + \beta_1 x_i$ is the center (median) of the distribution $\tilde{D}_{x_i}(\cdot)$.

It is apparent, that the probability integral transform does not work in this example as the u_is are not Uniform (0,1); as eq. (40) suggests their distribution is a point mass at 1/2. Nevertheless, they do have the *same* distribution, hence the model-free prediction still works giving perfect point predictions:

$$\tilde{D}^{-1}_{x_{\mathrm{f}}}(u_i) = \tilde{D}^{-1}_{x_{\mathrm{f}}}(1/2) = \beta_0 + \beta_1 x_{\mathrm{f}} \quad \text{for all} \quad i.$$

We now consider a more problematic model where $Y_t = \beta_0 + \beta_1 x_t + c_t \varepsilon_t$ where $x_t = t$ for $t = 1, ..., n$, $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$, and $c_t = \mathbf{1}\{t \geq n/2\}$. In other words, the first half of the scatterplot has no error like the previous example but the second half may have appreciable error; see Figure 7 (a) for an illustration. Here, we have $u_t \simeq 1/2$ for all $t < n/2$, but $u_t \sim$ i.i.d. Uniform (0,1) for $t \geq n/2$. This mixed quality of the transformed variables u_t causes the model-free prediction method to break down.

Fortunately, in both the above examples the problem can be diagnosed by an exploratory investigation of the transformed variables u_i much like the usual diagnostics on residuals in regression. It is obvious that non-uniformity of the u_is is a red flag, and can be easily diagnosed by a histogram and/or Q-Q plot. In particular, if the

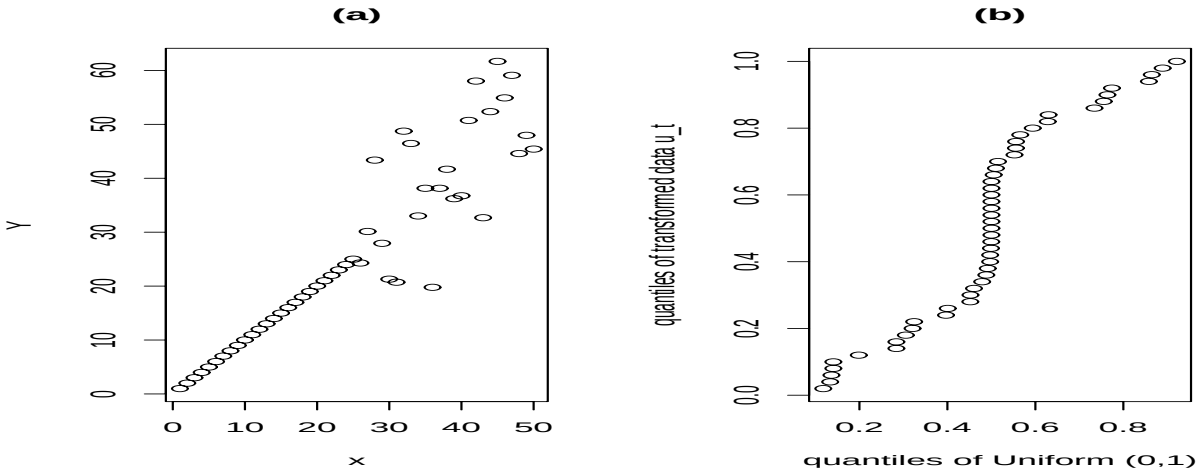

Figure 7: (a) Scatterplot of model $Y = 2x + \mathbf{1}\{x \geq 25\} \cdot \varepsilon_x$ for $x = 1, ..., 50$ with $\varepsilon_x \sim$ i.i.d. $N(0, 100)$. (b) Q-Q plot of the transformed variables u_t vs. the quantiles of Uniform (0,1).

distribution of the u_is appears to contain a point mass at 1/2 or elsewhere, then a problem is identified; for example, the Q-Q plot of Figure 7 (b) clearly indicates the presence of a point mass on 1/2.

Finally, let us go back to the homoscedastic example, where $Y = \beta_0 + \beta_1 x + \varepsilon_x$ with $\varepsilon_x \sim$ i.i.d. $N(0, \sigma^2)$ for $\sigma^2 > 0$. Even if σ^2 is very small, the situation can be salvaged from a model-free point of view by a careful design of the x points that would ensure $N_{x,h}$ is large for all x with h small enough that $h|\beta_1|$ is also small; if $|\beta_1|$ is appreciable, this would either require obtaining multiple Y responses associated with each design point x and/or employing a very high density of the x points to be used.

4.7 Simulation: when a nonparametric regression model is true

The building block for the simulation in this subsection is model (12) with $\mu(x) = \sin(x)$, $\sigma(x) = (\cos(x/2) + 2)/7$, and errors ε_t i.i.d. N(0,1) or two-sided exponential (Laplace) rescaled to unit variance. 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

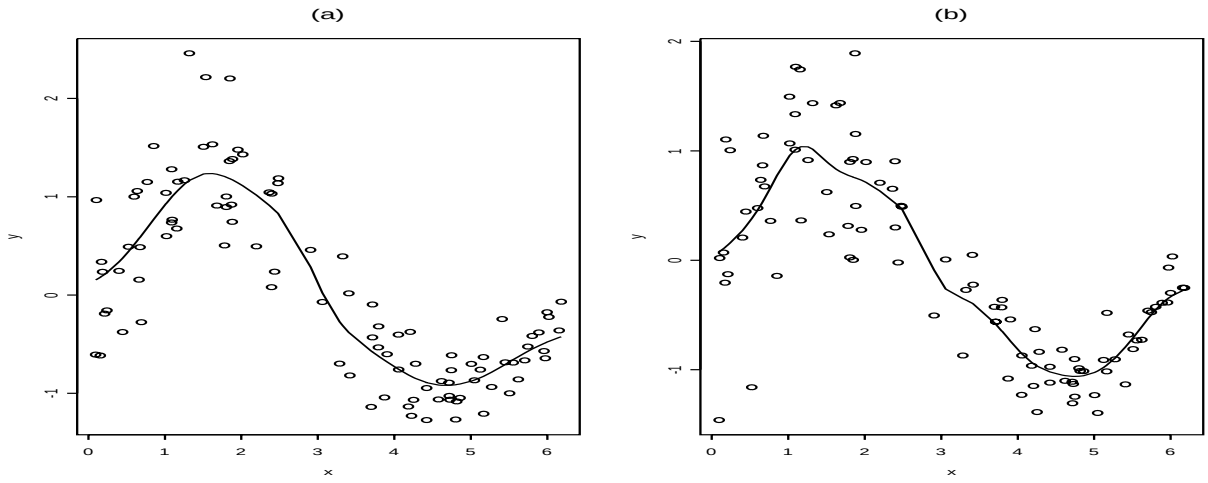

Figure 8: Typical scatterplots with superimposed kernel smoothers; (a) Normal data; (b) Laplace data.

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 (27). For all methods (except the NORMAL) the correction of Remark 4.7 was employed. The required bandwidths were computed by L_1 (PRESAR) cross-validation as described in Appendix B. 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.

For each type of interval, the corresponding empirical coverage level (COV) 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).

Attention focused on two possible prediction points, namely $x_{\text{f}} = \pi$ and $x_{\text{f}} = \pi/2$. The first point represents a case where $\mu(x)$ displays high slope but zero curvature; in the second case, the situation is reversed: zero slope but high curvature. The latter is actually a 'peak' of the function $\mu(x)$, and results into large bias of nonparametric estimators of $\mu(x)$. Note that the point $x_{\text{f}} = 3\pi/2$ corresponds to a 'valley' of the

function $\mu(x)$; the situation here is distributionally identical to that of the case $x_f = \pi/2$, and thus is omitted. Since $x_f = \pi/2$ and $x_f = \pi$ are extreme points in terms of curvature and bias, it is expected that points in-between would result in prediction interval performance that is somewhere in-between the relevant entries of Table 4.2 below.

As previously mentioned, in the practical construction of bootstrap predictive intervals one would employ a large number of bootstrap simulations, say $B = 1{,}000$ or 2,000. Nevertheless, bootstrap predictive intervals are very computer-intensive; hence, for the purposes of our simulation this number was curtailed to $B = 333$. Even with $B = 333$ and with the generation of just 500 series for each scenario, the compilation of the entries of Table 4.2 takes five days of CPU time on a standard 2.5GHz PC. Of course, simulations (including bootstrap) are especially amenable to parallel computing that can drastically reduce the computation time; the author took advantage of the Triton Resource at the San Diego Supercomputer Center of UCSD. The R functions used in the computation are provided (with absolutely no warranty!) at: `http://www.math.ucsd.edu/~politis/SOFT/MF3functions.R`.

Table 4.2 summarizes our findings, and contains a number of important features:

- The NORMAL intervals are characterized by under-coverage even when the true distribution is Normal. In particular, in the case $x_f = \pi/2$, the NORMAL interval's under-coverage is striking; the reason is the high bias of the kernel estimator at the points of a 'peak' or 'valley' that the normal interval (27) 'sweeps under the carpet'.

- The length of the NORMAL intervals is quite less variable[11] 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.

- The MF/MB intervals are more accurate than their MB analogs in the case $x_f = \pi/2$. However, in the case $x_f = \pi$, the MB intervals are most accurate, and the MF/MB intervals seem to *over*-correct (and over-cover); this over-coverage can be attributed to 'leakage' in the smoother bias and should be alleviated with a larger sample size or the use of undersmoothing—see the discussion below.

[11] The standard deviation of the length is estimated as 22.4 × st. err. where $22.4 \simeq \sqrt{500}$.

- Interestingly, the performance of MF^2 intervals resembles that of MB intervals; similarly, the performance of MF/MF^2 intervals resembles that of MF/MB intervals. As a matter of fact, the MF/MF^2 intervals have the best coverage in the case $x_\text{f} = \pi/2$; this is quite surprising since one would except the MB and MF/MB intervals to have a distinct advantage when model (12) is true.

- The price to pay for using the more generally valid MF/MF^2 intervals instead of the model-specific MF/MB ones here seems to be the increased variability associated with interval length of the former.

Distribution:	Normal		Laplace	
Case $x_\text{f} = \pi/2$	COV	LEN (st.err.)	COV	LEN (st.err.)
MB	0.760	0.992 (0.010)	0.788	0.986 (0.013)
MF/MB	0.838	1.260 (0.017)	0.836	1.211 (0.017)
MF^2	0.768	1.033 (0.011)	0.768	0.987 (0.015)
MF/MF^2	0.888	1.587 (0.022)	0.884	1.687 (0.027)
NORMAL	0.754	0.937 (0.004)	0.815	0.928 (0.004)
Case $x_\text{f} = \pi$				
MB	0.882	0.973 (0.010)	0.898	0.975 (0.010)
MF/MB	0.950	1.214 (0.012)	0.942	1.195 (0.013)
MF^2	0.884	0.989 (0.010)	0.888	0.985 (0.011)
MF/MF^2	0.970	1.510 (0.014)	0.954	1.584 (0.018)
NORMAL	0.877	0.937 (0.004)	0.874	0.933 (0.004)

Table 4.2. Empirical coverage levels (COV), and (average) lengths (LEN) of different prediction intervals with nominal coverage of 0.90; $n = 100$ and bandwidths chosen by L_1 cross-validation.

Finally, the problematic case $x_\text{f} = \pi$ 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 $x_\text{f} = \pi/2$ and $x_\text{f} = 3\pi/2$. 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 $x_\text{f} = \pi/2$ 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 that adversely influences the prediction performance at all points—even points associated with low estimation bias. This is the phenomenon previously referred to as *'bias leakage'*; it can be alleviated with a larger sample size and/or using higher-order smoothing kernels or other low bias approximation methods, e.g., wavelets. It can also be alleviated using bandwidth tricks such as *undersmoothing*—see the detailed discussion in Remark 3.2. 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.

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

In this subsection, we investigate the performance of the different prediction intervals in a set-up where model (12) is not true. For easy comparison with Section 4.7, we will keep the same (conditional) mean and variance, i.e., we will generate independent Y data such that $E(Y|x) = \sin(x)$, $Var(Y|x) = (\cos(x/2) + 2)/7$, and design points $x_1, \ldots, x_{100}$ equi-spaced on $(0, 2\pi)$ as before. However, the error structure $\varepsilon_x = (Y - E(Y|x))/\sqrt{Var(Y|x)}$ will be assumed to have to have skewness and/or kurtosis that depends on x, thereby violating the i.i.d. assumption.

So, for our simulation we will consider the simple construction:

$$\varepsilon_x = \frac{c_x Z + (1 - c_x)W}{\sqrt{c_x^2 + (1 - c_x)^2}} \tag{41}$$

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*.

Distribution of W:	χ^2_2		t_5	
Case $x_{\mathrm{f}} = \pi/2$	COV	LEN (st.err.)	COV	LEN (st.err.)
MB	0.768	0.948 (0.014)	0.762	0.972 (0.011)
MF/MB	0.844	1.230 (0.027)	0.844	1.206 (0.017)
MF^2	0.754	0.955 (0.015)	0.762	0.980 (0.013)
$\mathrm{MF/MF}^2$	0.880	1.646 (0.028)	0.882	1.616 (0.027)
NORMAL	0.843	0.930 (0.005)	0.801	0.937 (0.005)
Case $x_{\mathrm{f}} = \pi$				
MB	0.874	0.969 (0.010)	0.884	0.967 (0.010)
MF/MB	0.920	1.193 (0.012)	0.932	1.207 (0.011)
MF^2	0.878	0.968 (0.011)	0.862	0.988 (0.011)
$\mathrm{MF/MF}^2$	0.950	1.505 (0.016)	0.967	1.550 (0.017)
NORMAL	0.874	0.935 (0.005)	0.871	0.931 (0.005)
Case $x_{\mathrm{f}} = 3\pi/2$				
MB	0.744	0.484 (0.005)	0.766	0.491 (0.005)
MF/MB	0.836	0.618 (0.008)	0.850	0.607 (0.007)
MF^2	0.734	0.500 (0.006)	0.782	0.508 (0.006)
$\mathrm{MF/MF}^2$	0.902	0.745 (0.011)	0.910	0.738 (0.012)
NORMAL	0.980	0.928 (0.005)	0.978	0.939 (0.005)

Table 4.4. Entries as in Table 4.2 but with errors ε_x from eq. (41)

Table 4.4 presents our findings; they are qualitatively similar to those of Table 4.2 although differences between methods are more accentuated. In particular:

- The NORMAL intervals are totally unreliable which is to be expected due to the non-normal error distributions.
- The $\mathrm{MF/MF}^2$ intervals are the best (by far) in the cases $x_{\mathrm{f}} = \pi/2$ and $x_{\mathrm{f}} = 3\pi/2$ attaining close to nominal coverage even with a sample size as low as $n = 100$.
- The case $x_{\mathrm{f}} = \pi$ remains problematic for the same reasons previously discussed.

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'*, and its closely related Gaussian 'stepping-stone'. The idea behind those two principles is transforming the data into a domain that is easier to work with, e.g. an i.i.d. set-up or a Gaussian set-up respectively. The latter may be most useful for dependent data as it reduces the task of empirically assessing independence to the easier one of assessing uncorrelatedness. However, as demonstrated in Sections 3 and 4, the model-free prediction principle, i.e., the trasformation to an i.i.d. setting, 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 obliterate 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.

All in all, the paper presents a novel philosophy for statistical inference that encompasses standard methods such as Least Squares (see subsection 3.6) or nonparametric regression (see subsection 4.2).

Appendix A: the solution of eq. (20).

Squaring eq. (20) and using (21) we obtain the double solution:

$$Y_{\rm f} = \frac{m_{x_{\rm f}}(1-c)(1-c-cW_{\rm f}^2) \pm |W_{\rm f}|\sqrt{(1-c)^2 m_{x_{\rm f}}^2(-1+c+cW_{\rm f}^2)+(1-c)M_{x_{\rm f}}D_{\rm f}}}{D_{\rm f}} \tag{A.1}$$

where $s_{x_{\rm f}}^2 = M_{x_{\rm f}} - m_{x_{\rm f}}^2$, and $D_{\rm f} = (1-c)^2 + (c^2-c)W_{\rm f}^2$. A little algebra shows that the denominator $D_{\rm f}$ is strictly positive and the argument of the square root in eq. (A.1) is nonnegative provided the bound (A.2) below holds:[12]

$$|W_t| < \sqrt{\frac{1-c}{c}} \ \text{ for all } t. \tag{A.2}$$

To see that (A.2) is indeed true, note that eq. (18) implies

$$\frac{1}{W_t^2} = \frac{\tilde{s}_{x_t}^2}{(Y_t - \tilde{m}_{x_t})^2} = \frac{\tilde{M}_{x_t} - \tilde{m}_{x_t}^2}{(Y_t - \tilde{m}_{x_t})^2}$$

$$= \frac{cY_t^2 + (1-c)M_{x_t}^{(t)} - (cY_t + (1-c)m_{x_t}^{(t)})^2}{(1-c^2)(Y_t - m_{x_t}^{(t)})^2}$$

$$= \frac{c-c^2}{(1-c^2)} + \frac{(1-c)\left(M_{x_t}^{(t)} - (m_{x_t}^{(t)})^2\right)}{(1-c^2)(Y_t - m_{x_t}^{(t)})^2} \geq \frac{c-c^2}{(1-c^2)}$$

since[13] $M_{x_t}^{(t)} - (m_{x_t}^{(t)})^2 \geq 0$. From the above, it follows that $|W_t| \leq \sqrt{(1-c)/c}$ as desired, with *strict* inequality provided $M_{x_t}^{(t)} > (m_{x_t}^{(t)})^2$.

Now as previously noted, c is in general a small number. For example, if $c = K(0)/\sum_{k=1}^n K\left(\frac{x_t - x_k}{h}\right)$, then c tends to zero as $h \to 0$ in which case eq. (A.1) becomes

$$Y_{\rm f} \simeq m_{x_{\rm f}} \pm |W_{\rm f}| s_{x_{\rm f}}. \tag{A.3}$$

Comparing eq. (A.3) to eq. (20), it follows that the solution $Y_{\rm f} \simeq m_{x_{\rm f}} + W_{\rm f} s_{x_{\rm f}}$ is the *uniquely* correct one for eq. (A.3). By the same token (and due to the continuity in

[12] If $c = 0$, the bound (A.2) is trivial: $|W_t| < \infty$.

[13] To ensure that $M_{x_t}^{(t)} \geq (m_{x_t}^{(t)})^2$, the bandwidths h and q must be the same.

the variable c), the double solution (A.1) reduces to the *unique* solution of eq. (20) given by

$$Y_{\rm f} = \frac{m_{x_{\rm f}}(1-c)(1-c-cW_{\rm f}^2) + W_{\rm f}\sqrt{(1-c)^2 m_{x_{\rm f}}^2(-1+c+cW_{\rm f}^2) + (1-c)M_{x_{\rm f}}D_{\rm f}}}{D_{\rm f}} \tag{A.4}$$

that simplifies to eq. (22) as claimed.$\diamond$

Appendix B: L_1 vs. L_2 cross-validation.

Early proponents of (predictive) cross-validation include Allen (1971, 1974), Geisser (1971, 1975), and Stone (1974). Minimizing the PREdictive Sum of Squared residuals (PRESS) has been shown to be generally consistent for the optimal bandwidth—although characterized by slow rates of convergence; see e.g. Härdle and Marron (1985), and Härdle, Hall, and Marron (1988).

To further discuss the cross-validation procedure, we will focus here on the non-parametric model (12) with the objective of prediction of $Y_{\rm f}$ under the two criteria L_1 and L_2; see Table 3.1 for a summary. Since the L_2–optimal predictor is the one minimizing the Mean Squared Error (MSE) of prediction, the minimization of PRESS makes perfect sense in order to further reduce this MSE. However, the L_1–optimal predictor is the one minimizing the Mean Absolute Error (MAE) of prediction; to fine-tune it, it may be preferrable to use an L_1—cross-validation criterion, i.e., to minimize the PREdictive Sum of Absolute Residuals abbreviated as PRESAR $= \sum_{t=1}^{n} |\tilde{e}_t|$ where $\tilde{e}_t$ are the predictive residuals of eq. (17).

L_1—cross-validation may be advisable also on robustness considerations. Note that the random variable ε_t^2 (of which $\tilde{e}_t^2$ is a proxy) has a distribution with potentially heavy tails. For example, if $\varepsilon_t \sim N(0,1)$, then the density of ε_t^2 at point u has tails of type: $|u|^{-1/2}\exp(-|u|)$, i.e., tails of exponential thickness. If ε_t is itself a (two-sided) exponential, then the matters are much worse: the density of ε_t^2 at point u has tails of type: $|u|^{-1/2}\exp(-\sqrt{|u|})$. Now recall that $n^{-1}\times$PRESS $= n^{-1}\sum_{t=1}^{n}\tilde{e}_t^2$ is an empirical version of $E\varepsilon_t^2$. Although this expectation is finite in the two cases discussed above, the heavy tails of ε_t^2 make a sample average like $n^{-1}\times$PRESS somewhat unstable in practice. In other words, the presence of a large value generated by the heavy tails (or by potential outliers) can throw off PRESS together with the resulting bandwidths estimated by cross-validation. For this reason, L_1—cross-validation

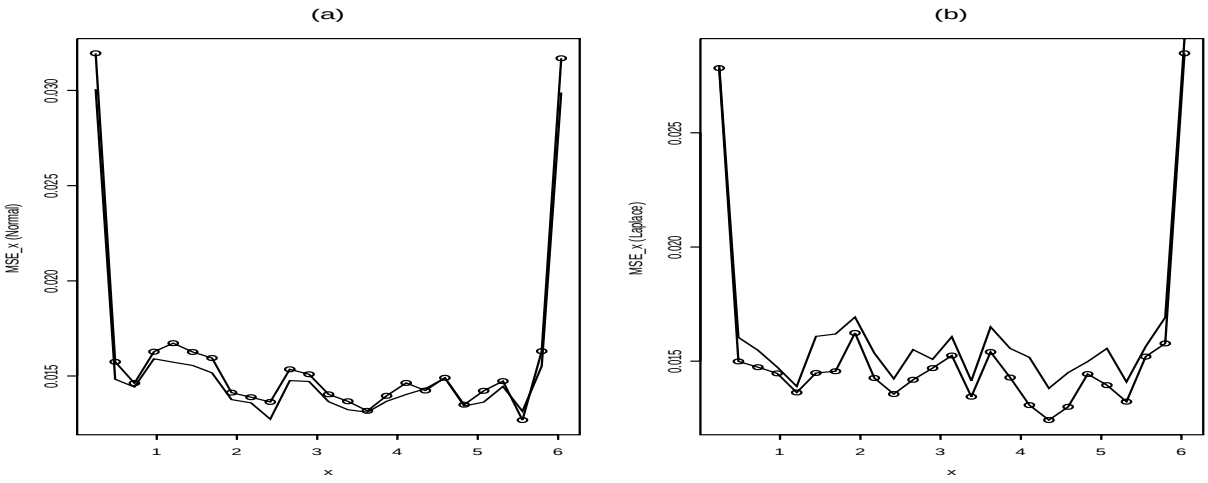

Figure 9: Plot of estimated MSE_x as a function of x in the case $\tau = 4$ using either L_1 (—o—) or L_2 cross-validation (——). (a) Normal data; (b) Laplace data.

may be preferable, and is not any more computationally expensive than the usual L_2—cross-validation.[14]

To see the difference between L_1 and L_2 cross-validation in practice, a small simulation was conducted. For the simulation, data were generated from model (12) with the choices $\mu(x) = \sin(x)$, $\sigma(x) = 1/10$, $\varepsilon_t \sim$ i.i.d. $(0, \tau^2)$ with distribution normal or two-sided exponential (Laplace), and different values for τ; reducing the error standard deviation τ has a similar effect as increasing sample size. For each of the error distributions, 999 datasets each of size $n = 100$ were created; the design points $x_1, \ldots, x_n$ were drawn each time from a uniform distribution on $(0, 2\pi)$.

The MSE of estimator m_x is denoted by MSE_x and was empirically evaluated at 25 different x–points taken equi-spaced on a grid of the interval $(0, 2\pi)$; those points were: $0.24, 0.48, \cdots, 5.79, 6.03$. Figure 9 shows a plot of the estimated MSE_x as a function of x in the case $\tau = 4$ using either L_1 or L_2 cross-validation. The peaking of the MSE at the boundaries is a well-known problem associated with kernel smoothers; it can be alleviated using the reflection technique of Hall and Wehrly (1991) which, in effect, makes the kernel estimator approximately equivalent to local linear fitting when the data are evenly distrubuted on the x-scale—see e.g. Fan and Gijbels (1996)

[14] In the rare case of non-unique minima in PRESAR cross-validation, the dilemma may be resolved by picking the result closest to one given by PRESS.

or Hastie and Loader (1993).

The performance of PRESS appears slightly better in the Normal case—see Figure 9(a), while PRESAR has a definite (and seemingly uniform) advantage in the Laplace case—see Figure 9(b). This is hardly surprising since minimization of $\sum_{t=1}^{n} \varepsilon_t^2$ (resp. $\sum_{t=1}^{n} |\varepsilon_t|$) is tantamount to Maximum Likelihood in the Normal (resp. Laplace) case. However, note that PRESAR's target is minimization of the Mean Absolute Error (MAE) of estimator m_x and *not* its MSE; the fact that PRESAR yields MSE's that are smaller than that from PRESS (whose target is MSE minimization) is quite noteworthy.

Estimating MSE_x on a grid of points also gives a natural estimate of the Integrated MSE of m_x denoted by $IMSE = \int_0^{2\pi} MSE_x \, dx$. Table B.1 compares the $IMSE$ of m_x using either L_1 or L_2 cross-validation for the bandwidth. The standard error of each entry of Table B.1 is approximately 0.01 as evaluated using subsampling; see e.g. Politis, Romano and Wolf (1999). The implication is that the two methods are very similar in the Gaussian case (with PRESS being slightly better); however, as expected, L_1 cross-validation has a definite advantage in the heavy-tailed case, and this is particularly true when the error variance is large (and/or the sample size is small).

$\tau =$	1	2	4
Normal	1.010	1.026	1.034
Laplace	0.970	0.959	0.941
Contam.	0.987	0.934	0.887

Table B.1. Entries are estimated ratios $IMSE(L_1)/IMSE(L_2)$ where L_1 and L_2 indicate the type of cross-validation used, and τ^2 is the error variance.

The simulation was repeated in a situation involving outliers; here the errors were $\varepsilon_t \sim$ i.i.d. $N(0, \tau^2)$ with a 5% contamination of $N(0, (10\tau)^2)$. Not surprisingly, PRESAR displays *robustness* to outliers and clearly outperforms PRESS in this case as indicated by the last row of Table B.1. As a consequence of the above discussion, it seems that PRESAR may be preferrable to PRESS overall since (a) it is optimal for the L_1 predictor, and (b) it works very well *even* for the L_2 predictor and MSE minimization—outperforming PRESS cross-validation in the non-normal examples.

Finally, note that if our ojective is prediction of $g(Y_{\mathrm{f}})$, then ideally our cross-

validation procedure would focus on the predictive residuals obtained from predicting $g(Y_t)$ on the basis of the delete-Y_t dataset; see Section 4.3 for more details.

References

[1] Allen, D.M. (1971). Mean square error of prediction as a criterion for selecting variables, *Technometrics*, 13, 469-475.

[2] Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, 16, 1307-1325.

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

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

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

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

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

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

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

[10] Caroll, R.J. and Ruppert, D. (1991). Prediction and tolerance intervals with transformation and/or weighting, *Technometrics*, 33, 197-210.

[11] 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.

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

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

[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] Geisser, S. (1971). The inferential use of predictive distributions. *Foundations of Statistical Inference*, (B.P. Godambe and D.S. Sprott, Eds.), Holt, Rinehart and Winston, Toronto, pp. 456-469.

[21] Geisser, S. (1975). The predictive sample re-use method with applications, *J. Amer. Statist. Assoc.*, 70, 320-328.

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

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

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

[25] 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.

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

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

[28] Härdle, W., Hall, P. and Marron, J.S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion), *J. Amer. Statist. Assoc.*, 83, 86-95.

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

[30] Härdle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation, *Ann. Statist.*, 13, 1465-1481.

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

[32] Hastie, T. and Loader, C. (1993). Local regression: automatic kernel carpentry, *Statist. Sci.*, vol. 8, no. 2, pp. 120-143.

[33] 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.

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

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

[36] Linton, O.B., Chen, R., Wang, N. and Härdle, W. (1997). An analysis of transformations for additive nonparametric regression, *J. Amer. Statist. Assoc.*, 92, 1512-1521.

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

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

[39] McCulloch, C.E. (2000). Generalized linear models. *J. Amer. Statist. Assoc.*, 95, 1320–1324.

[40] McMurry, T. and Politis, D. N. (2008). Bootstrap confidence intervals in nonparametric regression with built-in bias correction, *Statist. Prob. Letters,* vol. 78, no. 15, 2463–2469.

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

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

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

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

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

[46] 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.

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

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

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

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

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

[52] 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.

[53] Seber, G.A.F. and Lee, A.J. (2003). *Linear Regression Analysis, (2nd Ed.)*, Wiley, New York.

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

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

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

[57] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc., Ser. B*, 39, 144-147.

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

[59] 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.

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

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