Submitted to Statistical Science

# Models as Approximations II: A Model-Free Theory of Parametric Regression

**Andreas Buja**[*,‡], **Lawrence Brown**[*,‡], **Arun Kumar Kuchibhotla**[‡], **Richard Berk**[‡], **Ed George**[†,‡], **and Linda Zhao**[*,‡],

The Wharton School – University of Pennsylvania[‡]

*Abstract.* We develop a model-free theory of general types of parametric regression for iid observations. The theory replaces the parameters of parametric models with statistical functionals, to be called "regression functionals", defined on large non-parametric classes of joint $x$-$y$ distributions, without assuming a correct model. Parametric models are reduced to heuristics to suggest plausible objective functions. An example of a regression functional is the vector of slopes of linear equations fitted by OLS to largely arbitrary $x$-$y$ distributions, without assuming a linear model (see Part I). More generally, regression functionals can be defined by minimizing objective functions or solving estimating equations at joint $x$-$y$ distributions. In this framework it is possible to achieve the following: (1) define a notion of well-specification for regression functionals that replaces the notion of correct specification of models, (2) propose a well-specification diagnostic for regression functionals based on reweighting distributions and data, (3) decompose sampling variability of regression functionals into two sources, one due to the conditional response distribution and another due to the regressor distribution interacting with misspecification, both of order $N^{-1/2}$, (4) exhibit plug-in/sandwich estimators of standard error as limit cases of $x$-$y$ bootstrap estimators, and (5) provide theoretical heuristics to indicate that $x$-$y$ bootstrap standard errors may generally be more stable than sandwich estimators.

*AMS 2000 subject classifications:* Primary 62J05, 62J20, 62F40; secondary 62F35, 62A10.

*Key words and phrases:* Ancillarity of regressors, Misspecification, Econometrics, Sandwich estimator, Bootstrap, Bagging.

*Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 (e-mail: buja.at.wharton@gmail.com).*

[*]Supported in part by NSF Grant DMS-10-07657 and and DMS-1310795.

[†]Supported in part by NSF Grant DMS-14-06563.

*"The hallmark of good science is that it uses models and 'theory' but never believes them." (J.W. Tukey, 1962, citing Martin Wilk)*

## 1. INTRODUCTION

We develop in this second article a model-free theory of parametric regression, assuming for simplicity iid $x$-$y$ observations with quite arbitrary joint distributions. The starting point is the realization that regression models are approximations and should not be thought of as generative truths. A general recognition of this fact may be implied by the commonly used term "working model," but this vague term does not resolve substantive issues, created here by the fact that models are approximations and not truths. The primary issue is that model parameters define meaningful quantities only under conditions of model correctness. If the idea of models as approximations is taken seriously, one has to extend the notion of parameter from model distributions to basically arbitrary distributions. This is achieved by what is often called "projection onto the model," that is, finding for the actual data distribution the best approximating distribution within the model; one defines that distribution's parameter settings to be the target of estimation. Through such "projection" the parameters of a working model are extended to "statistical functionals," that is, mappings of largely arbitrary data distributions to numeric quantities. We have thus arrived at a functional point of view of regression, a view based on what we call ***regression functionals***.

The move from traditional regression parameters in correctly specified models to regression functionals obtained from best approximations may raise fears of opening the gates to irresponsible data analysis where misspecification is of no concern. No such thing is intended here. Instead, we rethink the essence of regression and develop a new notion of ***well-specification of regression functionals***, to replace the notion of *correct specification of regression models*. In the following bullets we outline an argument in the form of simple postulates.

- The essence of regression is the asymmetric analysis of association: Variables are divided into response and regressors.
- Interest focuses on properties of the conditional distribution of the response given the regressors.
- The goal or, rather, the hope is that the chosen quantities/functionals of interest are properties of the observed conditional response distribution, irrespective of the regressor distribution.
- Consequently, a regression functional will be called ***well-specified*** if it is a property of the observed conditional response distribution at hand, *irrespective of the regressor distribution.*

The first bullet is uncontroversial: asymmetric analysis is often natural, as in the contexts of prediction and causation. The second bullet remains at an intended level of vagueness as it explains the nature of the asymmetry, namely, the focus on the regressor-conditional response distribution. Intentionally there is no mention of regression models. The third bullet also steers clear of regression models by addressing instead quantities of interest, that is, regression functionals. In this and the last bullet, the operational requirement is that the quantities of interest not depend on the regressor distribution. It is this constancy across regressor

distributions that turns the quantities of interest into properties of the conditional response distribution alone.

All this can be made concrete with reference to the groundwork laid in Part I, Section 4. Consider the regression functional consisting of the vector of slopes obtained from OLS linear regression. It was shown in Part I that the OLS slope vector does not depend on the regressor distribution (is well-specified) if and only if the conditional response mean is a linear function of the regressors. Linearity, however, is the same as constancy of the gradient. Thus we have connected well-specification of the slope functional (its constancy across regressor distributions) to a property of the conditional response distribution (constancy of the gradient of the conditional response mean across the regressor space). This equivalence infuses the slope functional with special meaning on those conditional response distributions for which it is well-specified. Comparing well-specification of the slope functional with correct specification of the linear model, it is clear that the former is a weaker condition that sets aside the other linear model requirements (homoskedasticity, Gaussianity) not intimately tied to the slopes.

A desirable feature of the proposed definition of well-specification is that it generalizes to arbitrary types of parametric regression or, more precisely, to the statistical functionals derived from them. In particular, it applies to GLMs where the meaning of well-specified slopes is again correct specification of the mean function but setting aside other model requirements. Well-specification further applies to regression functionals derived from optimizing general objective functions or solving estimating equations. Well-specification finally applies to any ad hoc quantities if they define regression functionals for joint $x$-$y$ distributions.

The proposed notion of well-specification of regression functionals does not just define an ideal condition for populations but also lends itself to a tangible methodology for real data. A diagnostic for well-specification can be based on perturbation of the regressor distribution without affecting the conditional response distribution. Such perturbations can be constructed by reweighting the joint $x$-$y$ distribution with weight functions that only depend on the regressors. If a regression functional is not constant under such reweighting, it is misspecified.

In practice, use of this diagnostic often works out as follows. Some form of misspecification will be detected for some of the quantities of interest, but the diagnostic will also aid in interpreting the specifics of the misspecification. The reason is that reweighting essentially localizes the regression functionals. For the coefficients of OLS linear regression, for example, this means that reweighting reveals how the coefficients of the best fitting linear equation vary as the weight function moves across regressor space. Put this way, the diagnostic seems related to non-parametric regression, but its advantage is that it focuses on the quantities of interest at all times, while switching from parametric to non-parametric regression requires a rethinking of the meaning of the original quantities in terms of the non-parametric fit. To guide users of the diagnostic to insightful choices of weight functions, we introduce a set of specific reweighting methodologies, complete with basic statistical inference.

Following these methodological proposals, we return to the inferential issues raised in Part I and treat them in generality for all types of well-behaved regression functionals. We show that sampling variation of regression functionals has two sources, one due to the conditional response distribution, the other due

to the regressor distribution interacting with misspecification, where "misspecification" is meant in the sense of "violated well-specification" of the regression functional. A central limit theorem (CLT) shows that *both* sources, as a function of the sample size $N$, are of the usual order $N^{-1/2}$. Finally, it is shown that asymptotic plug-in/sandwich estimators of standard error are limits of $x$-$y$ bootstrap estimators, revealing the former to be an extreme case of the latter.

The present analysis becomes necessarily more opaque because algebra that worked out explicitly and lucidly for linear OLS in Part I is available in the general case only in the form of asymptotic approximation based on influence functions. Still, the analysis is now informed by the notion of well-specification of regression functionals, which gives the results a rather satisfactory form.

The article continues as follows. In Section 2 we discuss typical ways of defining regression functionals, including optimization of objective functions and estimating equations. In Section 3 we give the precise definition of well-specification and illustrate it with various examples. In Section 4 we introduce the reweighting diagnostic for well-specification, illustrated in Section 5 with specific reweighting methodologies applied to the LA homeless data (Part I). Section 6 shows for plug-in estimators of regression functionals how the sampling variability is canonically decomposed into contributions from the conditional response noise and from the randomness of the regressors. In Section 7 we state general CLTs analogous to the OLS versions of Part I. In Section 8 we analyze model-free estimators of standard error derived from the $M$-of-$N$ pairs bootstrap and asymptotic variance plug-in (often of the sandwich form). It holds in great generality that plug-in is the limiting case of bootstrap when $M \to \infty$. We also give some heuristics to suggest that boostrap estimators might generally be more stable than plug-in/sandwich estimators. In Section 9 we summarize the path taken in these two articles.

**Remark**: For notes on the history of model robustness, see Part I, Section 1. For the distinction between model robustness and outlier/heavy-tail robustness, see Part I, Section 13.

## 2. TARGETS OF ESTIMATION: REGRESSION FUNCTIONALS

This section describes some of the ways of constructing regression functionals, including those based on "working models" used as heuristics to suggest plausible objective functions. We use the following notations and assumptions throughout: At the population level there are two random variables, the regressor $\vec{\boldsymbol{X}}$ with values in a measurable space $\mathcal{X}$ and the response $Y$ with values in a measurable space $\mathcal{Y}$, with a joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$, a conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and a marginal regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. We express the connection between them using "$\otimes$" notation:

$$\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} \;=\; \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}. \tag{1}$$

Informally this is expressed in terms of densities by $p(y,\vec{\boldsymbol{x}}) = p(y|\vec{\boldsymbol{x}})p(\vec{\boldsymbol{x}})$. In contrast to Part I, the regressor and response spaces $\mathcal{X}$ and $\mathcal{Y}$ are now entirely arbitrary. The typographic distinction between $\vec{\boldsymbol{X}}$ and $Y$ is a hold-over from the OLS context of Part I. Both spaces, $\mathcal{X}$ *and* $\mathcal{Y}$, can be of any measurement type, univariate or multivariate, or even spaces of signals or images.

Regression functionals need to be defined on universes of joint distributions that are sufficiently rich to grant the manipulations that follow, including the

assumed existence of moments, influence functions, and closedness for certain mixtures. The details are tedious, hence deferred to Appendix A.1 without claim to technical completeness. The treatment is largely informal so as not to get bogged down in distracting detail. Also, the asymptotics will be traditional in the sense that $\mathcal{X}$ and $\mathcal{Y}$ are fixed and $N \to \infty$. For more modern technical work on related matters, see Kuchibhotla et al. (2018).

### 2.1 Regression Functionals from Optimization: ML and PS Functionals

In Part I we described the interpretation of slopes in linear OLS as regression functionals. The expression "linear OLS" is used on purpose to avoid the expression "linear models" because no model is assumed. Fitting a linear equation using OLS is a procedure to achieve a best fit of an equation by the OLS criterion. This approach can be generalized to other objective functions $\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$:

$$\boldsymbol{\theta}(\boldsymbol{P}) \;=\; \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \boldsymbol{E}_{\boldsymbol{P}}[\mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] \tag{2}$$

A common choice for $\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$ is the negative log-likelihood of a parametric regression model for $Y|\vec{\boldsymbol{X}}$, defined by a parametrized family of conditional response distributions $\{\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}\colon \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ with conditional densities $\{q(y\,|\,\vec{\boldsymbol{x}};\boldsymbol{\theta})\colon \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$. The model is not assumed to be correctly specified, and its only purpose is to serve as a heuristic to suggest an objective function:

$$\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) \;=\; -\log\, q(y\,|\,\vec{\boldsymbol{x}};\boldsymbol{\theta}). \tag{3}$$

In this case the regression functional resulting from (2) will be called a *ML functional*. It minimizes the Kullback-Leibler (KL) divergence of $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} = \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}$ and $\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}$, which is why one loosely interprets an ML functional as arising from a "projection of the actual data distribution onto the parametric model." ML functionals can be derived from major classes of regression models, including GLMs. Technically, they also comprise many M-estimators based on Huber $\rho$ functions (Huber 1964), including least absolute deviation (LAD, $L_1$) as an objective function for conditional medians, and tilted $L_1$ versions for arbitrary conditional quantiles, all of which can be interpreted as negative log-likelihoods of certain distributions, even if these may not usually be viable models for actual data. Not in the class of negative log-likelihoods are objective functions for M-estimators with redescending influence functions such as Tukey's biweight estimator (which also poses complications due to non-convexity).

Natural extensions of ML functionals can be based on so-called "proper scoring rules" (Appendix A.2) which arise as cross-entropy terms of Bregman divergences A special case is the expected negative log-likelihood arising as the cross-entropy term of KL divergence. The optimization criterion is the proper scoring rule applied to the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and model distributions $\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}$, averaged over regressor space with $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. The resulting regression functionals may be called "proper scoring functionals" or simply *PS functionals*, a superset of ML functionals. All PS functionals, including ML functionals, have the important property of Fisher consistency: If the model is correctly specified, i.e., if $\exists\, \boldsymbol{\theta}_0$ such that $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} = \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}_0}$, then the population minimizer is $\boldsymbol{\theta}_0$:

$$\text{if } \; \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} = \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}_0} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}, \;\; \text{then } \; \boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{\theta}_0. \tag{4}$$

See Appendix A.2 for background on proper scoring rules, Bregman divergences, and some of their robustness properties to outliers and heavy tailed distributions.

Further objective functions are obtained by adding parameter penalties to existing objective functions:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) \;=\; \mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) + \lambda \mathcal{R}(\boldsymbol{\theta}). \tag{5}$$

Special cases are ridge and lasso penalties. Note that (5) results in one-parameter families of penalized functionals $\boldsymbol{\theta}_\lambda(\boldsymbol{P})$ defined for populations as well, whereas in practice $\lambda = \lambda_N$ applies to finite $N$ with $\lambda_N \to 0$ as $N \to \infty$.

### 2.2 Regression Functionals from Estimating Equations: EE Funtionals

Objective functions are often minimized by solving stationarity conditions that amount to estimating equations with the scores $\boldsymbol{\psi}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) = -\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$:

$$\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] \;=\; \mathbf{0}. \tag{6}$$

One may generalize and define regression functionals as solutions in cases where $\boldsymbol{\psi}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$ is not the gradient of an objective function; in particular it need not be the score function of a negative log-likelihood. Functionals in this class will be called *EE functionals*. For OLS, the estimating equations are the normal equations, as the score function for the slopes is

$$\boldsymbol{\psi}_{OLS}(\boldsymbol{\beta}; y, \vec{\boldsymbol{x}}) \;=\; \vec{\boldsymbol{x}} y - \vec{\boldsymbol{x}} \vec{\boldsymbol{x}}' \boldsymbol{\beta} \;=\; \vec{\boldsymbol{x}}(y - \vec{\boldsymbol{x}}' \boldsymbol{\beta}). \tag{7}$$

A seminal work that inaugurated asymptotic theory for general estimating equations is by Huber (1967). A more modern and rigorous treatment is in Rieder (1994).

An extension is the "Generalized Method of Moments" (GMM, Hansen 1982). It applies when the number of moment conditions (the dimension of $\boldsymbol{\psi}$) is larger than the dimension of $\boldsymbol{\theta}$. An important application is to causal inference based on numerous instrumental variables.

Another extension is based on "Generalized Estimating Equations" (GEE, Liang and Zeger 1986). It applies to clustered data that have intra-cluster dependence, allowing misspecification of the variance and intra-cluster dependence.

### 2.3 The Point of View of Regression Functionals and its Implications

Theories of parametric models deal with the issue that a traditional model parameter has many possible estimators, as in the normal model $\mathcal{N}(\mu, \sigma^2)$ where the sample mean is in various ways the optimal estimate of $\mu$ whereas the median is a less efficient estimate of the same $\mu$. The comparison of estimates of the same traditional parameter has been proposed as a basis of misspecification tests (Hausman 1978) and called "test for parameter estimator inconsistency" (White 1982). In a framework based on regression functionals the situation presents itself differently. Empirical means and medians, for example, are not estimators of the same parameter; instead, they represent different statistical functionals. Similarly, slopes obtained by linear OLS and linear LAD are different regression functionals. Comparing them by forming differences creates new regression functionals that may be useful as diagnostic quantities, but in a model-robust framework there is no concept of "parameter inconsistency" (White 1982, p. 15), only a concept of differences between regression functionals.

A further point is that in a model-robust theory of observational (as opposed to causal) association, there is no concept of "omitted variables bias." There are only regressions with more or fewer regressor variables, none of which being "true" but some being more useful or insightful than others. Slopes in a larger regression are distinct from the slopes in a smaller regression. It is a source of conceptual confusion to write the slope of the $j$'th regressor as $\beta_j$, irrespective of what the other regressors are. In more careful notation one indexes slopes with the set of selected regressors $M$ as well, $\beta_{j\cdot M}$, as is done of necessity in work on post-selection inference (e.g., Berk et al. 2013). Thus the linear slopes $\beta_{j\cdot M}$ and $\beta_{j\cdot M'}$ for the $j$'th regressor, when it is contained in both of two regressor sets $M \neq M'$, should be considered as distinct regression functionals. The difference $\beta_{j\cdot M'} - \beta_{j\cdot M}$ is not a bias but a difference between two regression functionals. If it is zero, it indicates that the difference in adjustment between $M$ and $M'$ is immaterial for the $j$'th regressor. If $\beta_{j\cdot M'}$ and $\beta_{j\cdot M}$ are very different with opposite signs, there exists a case of Simpson's paradox for this regressor.

It should be noted that regression functionals generally depend on the full joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ of the response *and* the regressors. Conventional regression parameters describe the conditional response distribution only under correct specification, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} = \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}$, while the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is sidelined as ancillary. That the ancillarity argument for the regressors is not valid under misspecification was documented in Part I, Section 4. In the following sections this fact will be the basis of the notion of well-specification of regression functionals.

Finally, we state the following to avoid misunderstandings: In the present work, the objective is not to recommend particular regression functionals, but to point out the freedoms we have in choosing them and the conceptual clarifications we need when using them.

## 3. MIS-/WELL-SPECIFICATION OF REGRESSION FUNCTIONALS

The introduction motivated a notion of well-specification for regression functionals, and this section provides the technical notations. The heuristic idea is that a regression functional is well-specified for a joint distribution of the regressors and the response if it does not depend on the marginal regressor distribution. In concrete terms, this means that the functional does not depend on where the regressors happen to fall. The functional is therefore a property of the conditional response distribution alone.

### 3.1 Definition of Well-Specification for Regression Functionals

Recall the notation introduced in (1): $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} = \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}$. Here a technical detail requires clarification: conditional distributions are defined only almost surely with regard to $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, but we will assume that $\vec{\boldsymbol{x}} \mapsto \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}$ is a Markov kernel defined for all $\vec{\boldsymbol{x}} \in \mathcal{X}$.[1] With these conventions, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ uniquely determine $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} = \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}$ by (1), but not quite vice versa. Thus $\boldsymbol{\theta}(\cdot)$ can be written as

$$\boldsymbol{\theta}(\boldsymbol{P}) \;=\; \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}).$$

[1]Thus we assume a "regular version" has been chosen, as is always possible on Polish spaces.

**Definition:** *The regression functional $\boldsymbol{\theta}(\cdot)$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ if*

$$\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) \;=\; \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}'_{\vec{\boldsymbol{X}}})$$

*for all acceptable regressor distributions $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}'_{\vec{\boldsymbol{X}}}$.*

The term "acceptable" accounts for exclusions of regressor distributions such as those due to non-identifiability when fitting equations, in particular, perfect collinearity when fitting linear equations (see Appendix A.1).

**Remarks:**

- Importantly, the notion of well-specification is a *joint property* of a specific $\boldsymbol{\theta}(\cdot)$ and a specific $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. A regression functional will be well-specified for some conditional response distributions but not for others.
- The notion of well-specification represents an idealization, not a reality. Well-specification is never a fact, only degrees of misspecification are. Yet, idealizations are useful because they give precision and focus to an idea. Here, the idea is that a regression functional is intended to be a property of the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ alone, regardless of the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.

### 3.2 Well-Specification — Some Exercises and Special Cases

Before stating general propositions, here are some special cases to train intuitions.

- The OLS slope functional can be written $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{E}_{\boldsymbol{P}}[\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}_{\boldsymbol{P}}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})]$, where $\mu(\vec{\boldsymbol{x}}) = \boldsymbol{E}_{\boldsymbol{P}}[Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}]$. Thus $\boldsymbol{\beta}(\boldsymbol{P})$ depends on $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ only through the conditional mean function. The functional is well-specified if $\mu(\vec{\boldsymbol{x}}) = \boldsymbol{\beta}_0{}'\vec{\boldsymbol{x}}$ is linear, in which case $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}_0$. For the reverse, see Part I, Proposition 4.1.
- A special case is regression through the origin, which we generalize slightly as follows. Let $h(\vec{\boldsymbol{x}})$ and $g(y)$ be two non-vanishing real-valued square-integrable functions of the regressors and the response, respectively. Define

  $$\boldsymbol{\theta}_{h,g}(\boldsymbol{P}) \;=\; \frac{\boldsymbol{E}_{\boldsymbol{P}}[\,g(Y)h(\vec{\boldsymbol{X}})\,]}{\boldsymbol{E}_{\boldsymbol{P}}[\,h(\vec{\boldsymbol{X}})^2]}.$$

  Then $\boldsymbol{\theta}_{h,g}(\boldsymbol{P})$ is well-defined for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ if $\boldsymbol{E}_{\boldsymbol{P}}[\,g(Y)|\vec{\boldsymbol{X}}\,] = c\cdot h(\vec{\boldsymbol{X}})$ for some $c$.
- Ridge regression also defines a slope functional. Let $\boldsymbol{\Omega}$ be a symmetric non-negative definite matrix and $\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta}$ its quadratic penalty. Solving the penalized LS problem yields $\boldsymbol{\beta}(\boldsymbol{P}) = (\boldsymbol{E}_{\boldsymbol{P}}[\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}'] + \boldsymbol{\Omega})^{-1}\boldsymbol{E}_{\boldsymbol{P}}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})]$. This functional is well-specified if the conditional mean is linear, $\mu(\vec{\boldsymbol{x}}) = \boldsymbol{\beta}_0{}'\vec{\boldsymbol{x}}$ for some $\boldsymbol{\beta}_0$, and $\boldsymbol{\Omega} = c\boldsymbol{E}_{\boldsymbol{P}}[\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}']$ for some $c \geq 0$, in which case $\boldsymbol{\beta}(\boldsymbol{P}) = 1/(1+c)\,\boldsymbol{\beta}_0$, causing uniform shrinkage across all regression coefficients.
- Given a univariate response $Y$, what does it mean for the functional $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{E}_{\boldsymbol{P}}[\,Y]$ to be well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$? It looks as if it did not depend on the regressor distribution and is therefore always well-specified. This is a fallacy, however. Because $\boldsymbol{E}_{\boldsymbol{P}}[Y] = \boldsymbol{E}_{\boldsymbol{P}}[\mu(\vec{\boldsymbol{X}})]$, it follows that $\boldsymbol{E}_{\boldsymbol{P}}[Y]$ is independent of $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ iff the conditional response mean is constant: $\mu(\vec{\boldsymbol{X}}) = \boldsymbol{E}_{\boldsymbol{P}}[Y]$.

- Homoskedasticity: The average conditional variance functional $\sigma^2(\boldsymbol{P}) = \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{V}_{\boldsymbol{P}}[Y|\vec{\boldsymbol{X}}]]$ is well-specified iff $\boldsymbol{V}_{\boldsymbol{P}}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \sigma_0^2$ is constant, in which case $\sigma^2(\boldsymbol{P}) = \sigma_0^2$.
- The correlation coefficient $\rho(Y, X)$, if interpreted as a regression functional in a regression of $Y$ on $X$, is well-specified only in the trivial case when $\mu(X)$ is constant and $\boldsymbol{V}_{\boldsymbol{P}}[Y] > 0$, hence $\rho(Y, X) = 0$.
- Fitting a linear equation by minimizing least absolute deviations (LAD, the $L_1$ objective function) defines a regression functional that is well-specified if there exists $\boldsymbol{\beta}_0$ such that $\text{median}[\,\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}\,] = \boldsymbol{\beta}_0{}'\vec{\boldsymbol{X}}$.
- In a GLM regression with a univariate response and canonical link, the slope functional is given by

$$\boldsymbol{\beta}(\boldsymbol{P}) \;=\; \text{argmin}_{\boldsymbol{\beta}}\, \boldsymbol{E}_{\boldsymbol{P}}\big[\, b\,(\vec{\boldsymbol{X}}'\boldsymbol{\beta}) - Y\,\vec{\boldsymbol{X}}'\boldsymbol{\beta}\,\big],$$

  where $b(\theta)$ is a strictly convex function on the real line and $\theta = \vec{\boldsymbol{x}}'\boldsymbol{\beta}$ is the "canonical parameter" modeled by a linear function of the regressors. The stationary equations are[2]

$$\boldsymbol{E}_{\boldsymbol{P}}\big[Y\,\vec{\boldsymbol{X}}\big] \;=\; \boldsymbol{E}_{\boldsymbol{P}}\big[\,\partial b\,(\vec{\boldsymbol{X}}'\boldsymbol{\beta})\vec{\boldsymbol{X}}\,\big].$$

  This functional is well-specified iff $\boldsymbol{E}_{\boldsymbol{P}}\big[Y|\vec{\boldsymbol{X}}\big] = \partial b\,(\vec{\boldsymbol{X}}'\boldsymbol{\beta})$ for $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$. Well-specification of $\boldsymbol{\beta}(\boldsymbol{P})$ has generally no implication for $\boldsymbol{V}_{\boldsymbol{P}}\big[Y|\vec{\boldsymbol{X}}\big]$, except in the next example.
- Linear logistic regression functionals are a special case of GLM functionals where $Y \in \{0, 1\}$ and $b(\theta) = \log(1 + \exp(\theta))$. Well-specification holds iff $\boldsymbol{P}[\,Y = 1\,|\vec{\boldsymbol{X}}] = \phi(\vec{\boldsymbol{X}}'\boldsymbol{\beta})$ for $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ and $\phi(\theta) = \exp(\theta)/(1 + \exp(\theta))$. Because the conditional response distribution is Bernoulli, the conditional mean of $Y$ determines the conditional response distribution uniquely, hence well-specification of the regression functional $\boldsymbol{\beta}(\boldsymbol{P})$ is the same as correct specification of the logistic regression model.
- If $\boldsymbol{\theta}(\boldsymbol{P})$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, then so is the functional $f(\boldsymbol{\theta}(\boldsymbol{P}))$ for any function $f(\cdot)$. An example in linear regression is the predicted value $\boldsymbol{\beta}(\boldsymbol{P})'\vec{\boldsymbol{x}}$ at the regressor location $\vec{\boldsymbol{x}}$. Other examples are contrasts such as $\beta_1(\boldsymbol{P}) - \beta_2(\boldsymbol{P})$ where $\beta_j(\boldsymbol{P})$ denotes the $j$'th coordinate of $\boldsymbol{\beta}(\boldsymbol{P})$.
- A meaningless case of "misspecified functionals" arises when they do not depend on the conditional response distribution at all: $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}(\boldsymbol{P}_{\vec{\boldsymbol{X}}})$. Examples would be tabulations and summaries of individual regressor variables. They could not be well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ unless they are constants.

### 3.3 Well-Specification of ML, PS and EE Functionals

The following lemma, whose proof is obvious, applies to all ML functionals. The principle of pointwise optimization in regressor space covers also all PS functionals (see Appendix A.2.3, equation (14)).

[2] To avoid confusion with matrix transposition, we write $\partial b$ instead of $b'$ for derivatives.

**Proposition 3.3.1:** *If $\boldsymbol{\theta}_0$ minimizes $\boldsymbol{E}_{\boldsymbol{P}}[\mathcal{L}(Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}) \,|\, \vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}]$ for all $\vec{\boldsymbol{x}} \in \mathcal{X}$, then the minimizer $\boldsymbol{\theta}(\boldsymbol{P})$ of $\boldsymbol{E}_{\boldsymbol{P}}[\mathcal{L}(Y|\vec{\boldsymbol{X}};\boldsymbol{\theta})]$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, and $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}_0$ for all acceptable regressor distributions $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.*

The following fact is corollary of Proposition 3.3.1 but could have been gleaned from Fisher consistency (4).

**Proposition 3.3.2:** *If $\boldsymbol{\theta}(\cdot)$ is a ML or PS functional for the working model $\{\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, it is well-specified for all model distributions $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} = \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}$.*

The next fact states that an EE functional is well-specified for a conditional response distribution if it satisfies the EE conditionally and globally across regressor space for one value $\boldsymbol{\theta}_0$.

**Proposition 3.3.3:** *If $\boldsymbol{\theta}_0$ solves $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}_0; Y, \vec{\boldsymbol{X}})|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}] = \boldsymbol{0}$ for all $\vec{\boldsymbol{x}} \in \mathcal{X}$, then the EE functional defined by $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] = \boldsymbol{0}$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, and $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}_0$ for all acceptable regressor distributions $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.*

The proof is in Appendix A.4.

### 3.4 Well-Specification and Causality

The notion of well-specification for regression functionals relates to aspects of causal inference based on direct acyclic graphs (DAGs) and the Markovian structures they represent (e.g., Pearl (2009). Given a DAG, the theory explains which choices of regressors $\vec{\boldsymbol{X}}$ permit correct descriptions of causal effects for a given outcome variable $Y$. Focusing on one such choice of $\vec{\boldsymbol{X}}$ and $Y$, one is left with the task of describing interesting quantitative aspects of the conditional distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, which is thought to be unchanging under different manipulations and/or sampling schemes of the regressors $\vec{\boldsymbol{X}}$. Therefore, if a quantity of interest is to describe causal effects properly, it should do so irrespective of where the values of the causal variables $\vec{\boldsymbol{X}}$ have fallen. This is exactly the requirement of well-specification for regression functionals. In summary, well-specification of the quantities of interest is necessary for describing causal effects in DAGs.

Recently, Peters, Bühlmann and Meinshausen (2016, Section 1.1) discussed a related notion of "invariance" which can be interpreted as "invariance to regressor distributions". They propose this notion as a heuristic for causal discovery and inference based on multiple data sources with the same variables, one variable being singled out as the response $Y$. These multiple data sources are leveraged as follows: If for a subset of variables, $\vec{\boldsymbol{X}}$, the association $\vec{\boldsymbol{X}} \to Y$ is causal, then the conditional distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ will be the same across data sources. Subsets of causal variables $\vec{\boldsymbol{X}}$ with shared $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ across sources may therefore be discoverable if the sources differ in their regressor distributions and/or interventions on causal variables. For concreteness, the authors focus on a linear structural equation model (SEM), which allows us to reinterpret their proposals by abandoning the SEM assumption and consider instead the regression functional consisting of the OLS regression coefficients resulting from the linear SEM. Thus the proposed method is at heart an approach to detecting and inferring well-specified quantities, cast in a causal framework.

In the following section we will introduce a diagnostic for well-specification that can be interpreted as emulating multiple data sources from a single data source. The proposal is to systematically reweight the data to synthetically create alternative datasets. Peters et al. (2016, Section 3.3) briefly mention the idea of conditioning as related to the idea of multiple data sources. Such conditioning is naturally achieved by locally reweighting the data, as will be shown next.

## 4. A REWEIGHTING DIAGNOSTIC FOR WELL-SPECIFICATION: TARGETS AND INFERENTIAL TOOLS

Well-specification of regression functionals connects naturally to reweighting, both of populations and of data. A concrete illustration of the basic idea can be given by again drawing on the example of linear OLS: The OLS slope functional is well-specified iff $\boldsymbol{E}_{\boldsymbol{P}}[Y|\vec{\boldsymbol{X}}] = \boldsymbol{\beta}_0'\vec{\boldsymbol{X}}$ for some $\boldsymbol{\beta}_0$, in which case for any non-negative weight function $w(\vec{\boldsymbol{x}})$ we have $\boldsymbol{\beta}_0 = \operatorname{argmin}_{\boldsymbol{\beta}} \boldsymbol{E}_{\boldsymbol{P}}[w(\vec{\boldsymbol{X}})\,(Y - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2]$. Therefore the reweighting of interest is with regard to weights that are functions of the regressors only. The general reason is that such weights affect the distribution of the regressors but not the conditional response distribution. Reweighting provides an intuitive basis for diagnosing well-specification of regression functionals. Because of the practical importance of the proposed reweighting diagnostic, we insert this material early, deferring estimation and inference to Section 6.

Reweighting has an extensive history in statistics, too rich to recount. The present purpose of reweighting is methodological: to diagnose the degree to which the null hypothesis of well-specification of a regression functional is violated. To this end we propose what we call a "tilt test." It provides evidence of whether a real-valued regression functional is likely to rise or fall (tilt up or down) from one extreme of reweighting to another. The conclusions from a rejection based on this test are simple and interpretable.

In practice, the majority of regression functionals of interest are regression slopes connected to specific regressors. A more interesting problem than detection of misspecification is another question: Does misspecification impinge on the statistical significance of a slope of interest? That is, would a slope have lost or gained statistical significance if the regressor distribution had been different? This is the primary question to be addressed by the reweighting diagnostic.

### 4.1 Reweighting and Well-Specification

Consider reweighted versions of the joint distribution $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ with weight functions $w(\vec{\boldsymbol{x}})$ that depend only on the regressors, not the response, written as

$$\boldsymbol{P}^w_{Y,\vec{\boldsymbol{X}}}(dy, d\vec{\boldsymbol{x}}) = w(\vec{\boldsymbol{x}})\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}(dy, d\vec{\boldsymbol{x}}), \quad \text{or} \quad p^w(y, \vec{\boldsymbol{x}}) = w(\vec{\boldsymbol{x}})\, p(y, \vec{\boldsymbol{x}}),$$

where $w(\vec{\boldsymbol{x}}) > 0$ and $\boldsymbol{E}_{\boldsymbol{P}}[w(\vec{\boldsymbol{X}})] = 1$, which turns $\boldsymbol{P}^w_{Y,\vec{\boldsymbol{X}}}$ into a joint probability distribution for $(Y, \vec{\boldsymbol{X}})$ with the same support as $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$. At times, for specific weight functions, we will write $w(\vec{\boldsymbol{X}})\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ instead of $\boldsymbol{P}^w_{Y,\vec{\boldsymbol{X}}}$.

**Lemma 4.1**: $\boldsymbol{P}^w_{Y|\vec{\boldsymbol{X}}} = \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}^w_{\vec{\boldsymbol{X}}} = w(\vec{\boldsymbol{X}})\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.

The proof is elementary and simplest in terms of densities:

$$p^w(\vec{\boldsymbol{x}}) = \int p^w(y,\vec{\boldsymbol{x}})dy = \int w(\vec{\boldsymbol{x}})p(y,\vec{\boldsymbol{x}})dy = w(\vec{\boldsymbol{x}})\int p(y,\vec{\boldsymbol{x}})dy = w(\vec{\boldsymbol{x}})p(\vec{\boldsymbol{x}}),$$
$$p^w(y|\vec{\boldsymbol{x}}) = p^w(y,\vec{\boldsymbol{x}})/p^w(\vec{\boldsymbol{x}}) = (w(\vec{\boldsymbol{x}})\,p(y,\vec{\boldsymbol{x}}))/(w(\vec{\boldsymbol{x}})\,p(\vec{\boldsymbol{x}}) = p(y,\vec{\boldsymbol{x}})/p(\vec{\boldsymbol{x}}) = p(y|\vec{\boldsymbol{x}}).$$

We obtain as an immediate consequence:

**Proposition 4.1**: *If the regression functional $\boldsymbol{\theta}(\cdot)$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, it is unchanged under arbitrary $\vec{\boldsymbol{X}}$-dependent reweighting: $\boldsymbol{\theta}(\boldsymbol{P}^w_{Y,\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}(\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}})$.*

**Remark:** In fixed-$\boldsymbol{X}$ linear models theory, which assumes correct specification, it is known that reweighting the data with fixed weights grants unbiased estimation of coefficients. Translated to the current framework, this fact returns as a statement of invariance of well-specified functionals under $\vec{\boldsymbol{X}}$-dependent reweighting.

Tests of misspecification based on reweighting were proposed by White (1980a, Section 4) for linear OLS. The approach generalizes to arbitrary types of regression and regression functionals as follows: Given a weight function $w(\vec{\boldsymbol{X}})$ normalized for $\boldsymbol{P}$, the null hypothesis is $H_0 : \boldsymbol{\theta}(\boldsymbol{P}^w) = \boldsymbol{\theta}(\boldsymbol{P})$. For the case that $\boldsymbol{\theta}(\cdot)$ is the vector of OLS linear regression coefficients, White (ibid., Theorem 4) proposes a test statistic based on plug-in estimates and shows its asymptotic null distribution to be $\chi^2$. The result is a Hausman test (1978) whereby (using model-oriented language) an efficient estimate under the model is compared to an inefficient but consistent estimate. Rejection indicates misspecification. We will not draw on White's results but instead use the $x$-$y$ bootstrap as a basis of inference because (1) it directly applies to general types of regression under mild technical conditions, and (2) it lends itself to augmentation of visual displays that provide more informative diagnostics than vanilla tests. White (1980a) did not develop a methodology for reweighting tests other than recommending experimentation with multiple weight functions. The present goal is to introduce highly interpretable *one-parameter families of weight functions* and to illustrate their practical use to gain insights into the nature of misspecifications.

### 4.2 The Well-Specification Diagnostic: Population Version

In order to construct meaningful and interpretable weight functions, we construct them as functions of a *univariate variable* $Z$. This variable will often be one of the real-valued regressors, $Z = X_j$. However, the possibilities are more general: $Z$ may be a regressor not used by the regression functional, as when submodel slopes $\beta_{j\cdot M}$ (Section 2.3) are examined as to their well-specification with regard to weights that are a function of regressors not in the submodel (e.g., $Z = X_k$ where $k \notin M$). The variable $Z$ may also be a more general function of the regressors, $Z = f(\vec{\boldsymbol{X}})$, as when $Z = \boldsymbol{\beta}'\vec{\boldsymbol{X}}$ is the OLS fit of $Y$, or $Z = X_{j\bullet}$ is $X_j$ adjusted for all other regressors (Part I, Section 9).

Given a variable $Z$, consider for concreteness a univariate Gaussian weight function of $Z$, centered at $\xi$ on the $Z$ axis:

$$w_\xi(z) \;=\; w^*_\xi(z)/\boldsymbol{E}[w^*_\xi(Z)]\,, \qquad w^*_\xi(z) \;\propto\; \exp(-(z-\xi)^2/(2\sigma^2))\,, \tag{8}$$

where $\sigma$ is a user-specified bandwidth parameter (see Section 4.3 below).

Next consider a one-dimensional regression functional $\theta(\boldsymbol{P})$, such as a linear regression slope. A graphical diagnostic is obtained by plotting $\theta(\cdot)$ as a function of the reweighting centers $\xi$:

$$\xi \;\mapsto\; \theta_\xi(\boldsymbol{P}) \;=\; \theta(w_\xi(Z)\,\boldsymbol{P})\,. \tag{9}$$

If the regression functional $\theta(\boldsymbol{P})$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, then $\theta_\xi(\boldsymbol{P})$ is constant in $\xi$ and equal to $\theta(\boldsymbol{P})$. Equivalently, if $\theta_\xi(\boldsymbol{P})$ is not constant in $\xi$, then $\theta(\boldsymbol{P})$ is misspecified. Thus non-constancy is a sufficient criterion for misspecification. Insightful choices of traces of the form (9) will be proposed below.

### 4.3 The Reweighting Diagnostic: Data Version

To make the diagnostic actionable on data, one obtains estimates

$$\hat{\theta}_\xi \;=\; \theta(\hat{w}_\xi(Z)\widehat{\boldsymbol{P}}),$$

where $\hat{w}_\xi(x)$ is a weight function that is empirically normalized to unit mass, $\widehat{\boldsymbol{E}}[\hat{w}_\xi(Z)]=1$, where $\widehat{\boldsymbol{E}}[...]$ denotes the sample average. This means using weights for the observations of the form

$$w_i = \hat{w}_\xi(z_i) \propto \exp((z_i-\xi)^2)/(2(\alpha\hat{\sigma})^2), \quad \tfrac{1}{N}\textstyle\sum_i w_i = 1, \quad i=1,...,N.$$

We parametrize the bandwidth in terms of the empirical standard deviation $\hat{\sigma}$ of $Z$, and a multiplier $\alpha$. In the examples we use $\alpha=1$.

In order to plot a discretized version of the trace $\xi \mapsto \hat{\theta}_\xi$, we obtain estimates $\hat{\theta}_\xi$ for a grid of values $\xi_{(1)} < ... < \xi_{(K)}$ on the $Z$ axis, a simple choice being the interior deciles of the empirical $Z$ distribution. Hence $K=9$, unless $Z$ has numerous ties, causing some deciles to collapse. Finally, we plot $\xi_{(k)} \mapsto \hat{\theta}_{\xi_{(k)}}$. This is carried out in Figures 1-3 for the LA homeless data (see Section 5).

### 4.4 Interpretations of the Reweighting Diagnostic

The reweighting diagnostic is likely to be accessible to practitioners of regression. One reason is that the restriction to weights as a function of a univariate variable $Z$ permits a simple left-to-right comparison: Is $\xi \mapsto \theta(w_\xi(Z)\,\boldsymbol{P})$ higher or lower on the right than on the left? In our experience, the dominant feature of such traces is indeed monotonicity. The intuitive appeal of reweighting is further helped by two mutually compatible interpretations:

- **Data frequency**: Reweighting mimics scenarios of datasets that contain more or fewer observations as a function of $Z$ than the observed dataset. Thus it answers questions such as "what if there were more observations with low (or high) values of $Z$?"
- **Conditioning**: Reweighting can be seen as "soft conditioning on $Z$" in the sense that conditioning on "sharp inclusion" in an interval $\xi-c<Z<\xi+c$ is replaced by "soft inclusion" according to the weight function $w_\xi(z)$.

In what follows we use either of these interpretations depending on the context.

### 4.5 Inferential Features for Reweighting Diagnostics

Graphical diagnostics need inferential augmentation to answer questions of whether visually detected features are real. Presently the two main questions are:

(1) Is the variation/non-constancy in $\xi_{(k)} \mapsto \hat{\theta}_{\xi_{(k)}}$ sufficiently strong to be statistically significant and hence suggest misspecification of $\theta(\cdot)$?
(2) Where are the estimates $\hat{\theta}_{\xi_{(k)}}$ statistically significantly different from zero?

For regression slopes question (2) may be more relevant than (1) because one usually cares about their statistical significance. Therefore, to answer question (2), we decorate the diagnostic plot with traces of bootstrapped estimates, as shown in the plots of Figures 1-3. Bootstrap resampling is done from the actual, not the reweighted, data. The weight functions have the same centers $\xi_{(k)}$, but their bandwidth is based on bootstrapped standard deviations. In the figures we show 199 bootstrap traces in gray color, amounting to a so-called "spaghetti plot". Along with the bootstrap replications we also show bootstrap error bars at the grid locations. Their widths are $\pm 2$ bootstrap standard errors.

As can be illustrated with Figures 1-3, statistical significance can feature a variety of patterns. Significance may exist ...

(2a) ... across the whole range of reweighting centers $\xi_{(k)}$ and in the same direction, as in the top right plot of Figure 1;
(2b) ... both on the left and the right but in opposite directions with a transition through insignificance in between, as is nearly the case in the center left plot of Figure 2;
(2c) ... over part of the range, typically the left or the right side; such tendencies are seen in the two center plots of Figure 1;
(2d) ... nowhere, as in the bottom right plot of Figure 2.

To answer question (1) regarding the presence of misspecification, we piggyback on the bootstrap exercise meant to answer question (2). Because most detections of misspecification arise from a monotone tilt in the trace $\xi_{(k)} \mapsto \hat{\theta}_{\xi_{(k)}}$, we construct a cheap test statistic by forming the difference between the two extreme points of the trace, $\hat{\theta}_{\xi_{(K)}} - \hat{\theta}_{\xi_{(1)}}$.[3] We obtain its bootstrap distribution almost for free, hence we can perform a crude bootstrap test by placing the null value zero in the bootstrap distribution. The bootstrap p-value and the test statistic are shown near the top of each plot frame in Figures 1-3. For example, the top left frame of Figure 1 shows "`Tilt: p=0.04 d=2.18`", meaning that the difference of 2.18 is statistically significant with a (two-sided) p-value of 0.04.[4]

Finally we show on the left side of each frame a visual version of unweighted plain statistical significance of the quantity of interest in the form of a bootstrap confidence interval around the unweighted estimate $\hat{\theta} \pm 2$ unweighted bootstrap standard errors. In addition, we show 199 bootstrap estimates (gray points horizontally jittered to reduce overplotting). The location on the horizontal axis has no meaning other than being 10% to the left of the range $(\xi_{(1)}, \xi_{(K)})$ of the traces.

## 5. THE REWEIGHTING DIAGNOSTIC FOR WELL-SPECIFICATION: METHODOOOGY AND EXAMPLES

The following subsections demonstrate three different purposes of the diagnostic. The quantities of interest are linear OLS slopes, though the approach generalizes to many types of regression:

[3] This test statistic does not result in a Hausman (1978) test: both estimates are "inefficient under correct model specification." However, it quantifies an obvious visual feature of the traces.
[4] For 199 bootstrap replicates the lowest possible two-sided p-value is $0.01 = 2 \cdot 1/(1 + 199)$.

- **Focal Regressor**: Expose a regression slope $\beta_k(\boldsymbol{P})$ of special interest to reweighting on each regressor in turn: $Z = X_j$ for $j = 1, ..., p$ (Section 5.1). This produces highly interpretable insights into interactions of regressor $X_k$ with all other regressors $X_j$, without modeling these interactions directly.
- **Nonlinearity detection**: Expose each regression slope $\beta_j(\boldsymbol{P})$ to reweighting on its own regressor, $Z = X_j$ (Section 5.2). This produces insights into marginal nonlinear behaviors of response surfaces.
- **Focal Reweighting**: Use a single reweighting variable of interest (here: $Z = \boldsymbol{\beta}'\vec{\boldsymbol{X}}$) to diagnose well-specification of all slopes $\beta_j(\boldsymbol{P})$ (Section 5.3).

These diagnostics will be illustrated with the LA homeless data of Part I, Section 2. The observations consist of a sample of 505 census tracts in the LA metropolitan area, and the variables are seven quantitative measures of the tracts with largely self-explanatory names: The response is the `StreetTotal` (count) of homeless people in a census tract, and the six regressors are: `MedianIncome` (of households, in $1,000s), `PercMinority`, and the prevalences of four types of lots: `PercCommercial`, `PercVacant`, `PercResidential` and `PercIndustrial`.

### 5.1 Diagnostics for a Focal Regression Coefficient of Interest (Figure 1)

One variable stands out as potentially accessible to intervention by public policies: `PercVacant`. Vacant lots could be turned into playgrounds, sports fields, parks, or offered as neighborhood gardens.[5] It would therefore be of interest to check whether the regression coefficient of `PercVacant` possibly measures a causal effect, for which it is a necessary condition that it be well-specified (Section 3.4). To this end, Figure 1 shows diagnostics for the coefficient of `PercVacant` under reweighting on all six regressors.

As the plots show, statistical significance of the coefficient of `PercVacant` holds by and large under reweighting across the ranges of all six regressors. While this is comforting, there exists a weakening of significance in the extremes of the ranges of three regressors: high `MedianInc`, low `PercMinority` and low `PercResidential`. With these qualitative observations it is already indicated that well-specification of the coefficient of `PercVacant` is doubtful, and indeed the tilt tests show statistical significance with 2-sided p-values of 0.01 and 0.02 for `PercMinority` and `MedianInc`, respectively. The variable `PercResidential` also looks rather steep, but its tilt test has a weaker p-value around 0.1. Finally, a very weak indication is shown for larger effects at higher levels of `PercVacant`.

Does this indication of misspecification invalidate a causal effect of `PercVacant`? It does not. It only points to the likely possibility that the causal effect is not correctly described by a single linear regression coefficient; it is rather a more complex function of the regressors. Useful insight into the nature of the causal effect (if this is what it is) can be gleaned from the diagnostic plots by using them to answer an obvious question: Where is the effect of `PercVacant` likely to be strong? An answer might indeed help in prioritizing interventions. Interpreting the plots of Figure 1 liberally, one could state that the effect of `PercVacant` looks strongest for census tracts with high `PercMinority`, followed by high `PercResidential` and low `MedianInc`. These observations seem rather plausible and may indeed point to census tracts worth prioritizing for intervention with public policies.

[5]Such programs have indeed been enacted in some cities. We abstain from commenting on the controversies surrounding such policies.

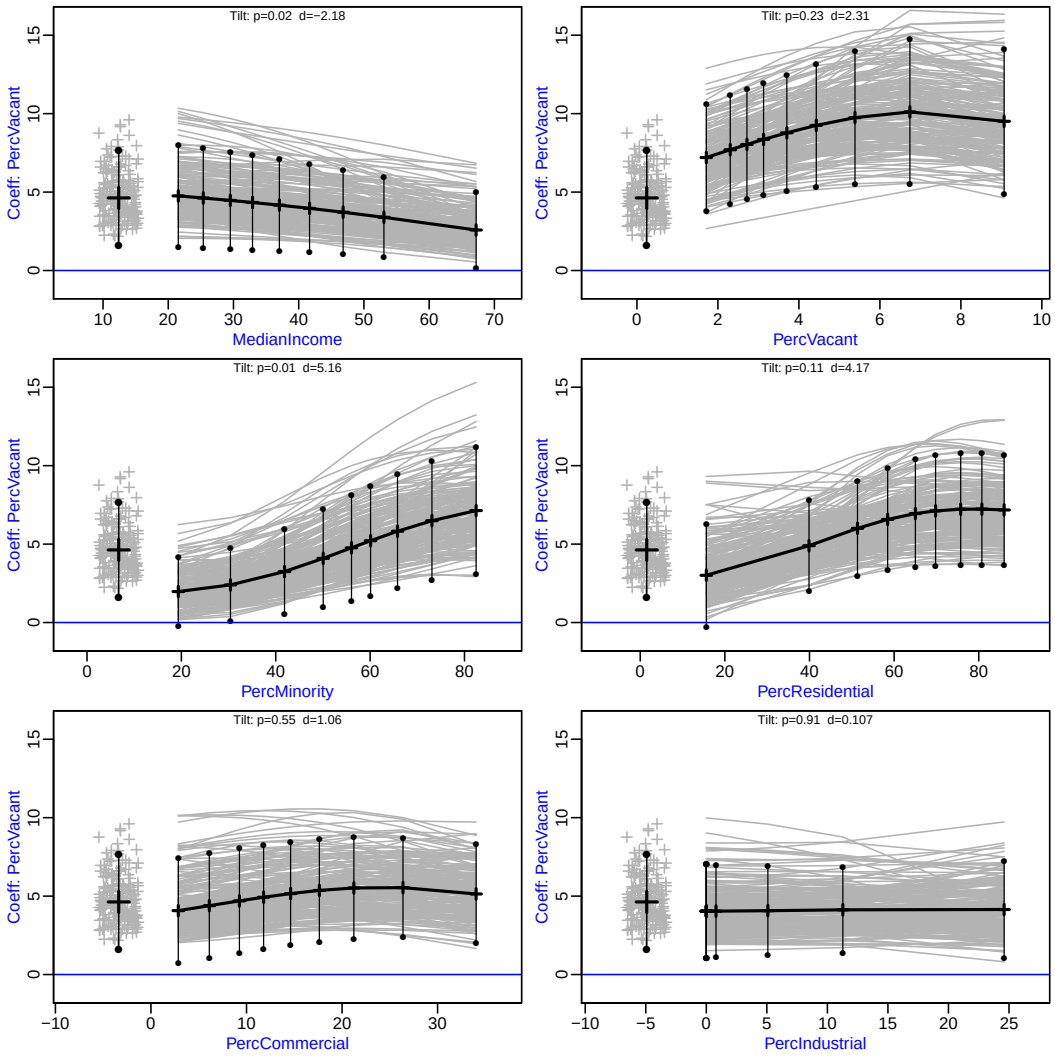


FIG 1. *Diagnostics for the slope of* `PercVacant`*; LA Homeless Data (see Section 2, Part I). Vertical axis = regression coefficient of* `PercVacant` *(all frames); horizontal axes = regressors.*

The insights gained so far point to the presence of interactions between `PercVacant` and other regressors because the slope of `PercVacant` varies at different levels of those other regressors. A natural next step would be more detailed modeling that includes interactions between `PercVacant` and the three interacting regressors, but the essential insights have already been gained.

## 5.2 Diagnostics for Slopes Reweighted by Their Own Regressors (Figure 2)

The top right plot in Figure 1 is a special case where the slope of interest is reweighted by its own regressor, `PercVacant`. It has a different interpretation, not related to interactions but to nonlinear effects. To get a better picture of the possibilities that can arise in real data, we show in Figure 2 the corresponding plots for all six regressors and their slopes.

Glancing at the six plots, we note some unpredictable effects of reweighting, both on the values and the estimation uncertainties of the slopes. We find examples of larger and smaller estimates as well as stronger and weaker statistical

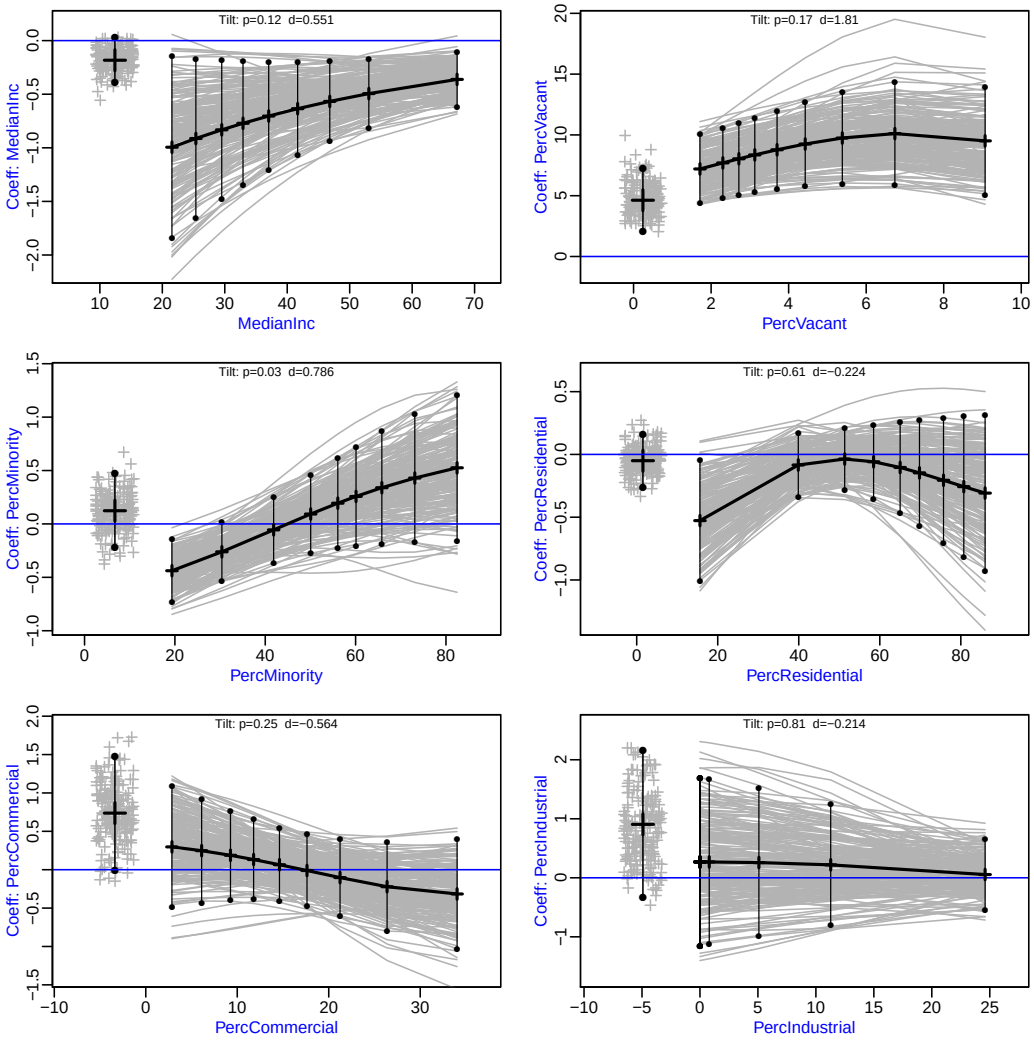


FIG 2. *Misspecification Diagnostics: Each slope reweighted by its own regressor — indications of nonlinearity.*

significances relative to their unweighted analogs:

- Bottom left plot for the regressor `PercCommercial`: The unweighted estimate of $\beta_j(\boldsymbol{P})$ (on the left side of the plot) is weakly statistically significant (the lower end of the $\pm 2$ standard error confidence interval touches zero). The reweighted estimates of $\beta_j(w_\xi(X_j)\boldsymbol{P})$, however, are closer to zero and nowhere statistically significant for any $\xi$ in the range of `PercCommercial`.
- Top right plot for the regressor `PercVacant`: The unweighted estimate and the reweighted estimates are all statistically significant, but the reweighted ones are systematically larger and much more statistically significant.

Another noteworthy case of a different nature appears for the regressor `PercMinority` (Figure 2, center left plot). While the unweighted estimate is statistically insignificant, the locally reweighted estimates reveal a striking pattern:

- For low values of `PercMinority` $\approx 20\%$, the slope is negative and statistically significant: Incrementally more minorities is associated with a lower

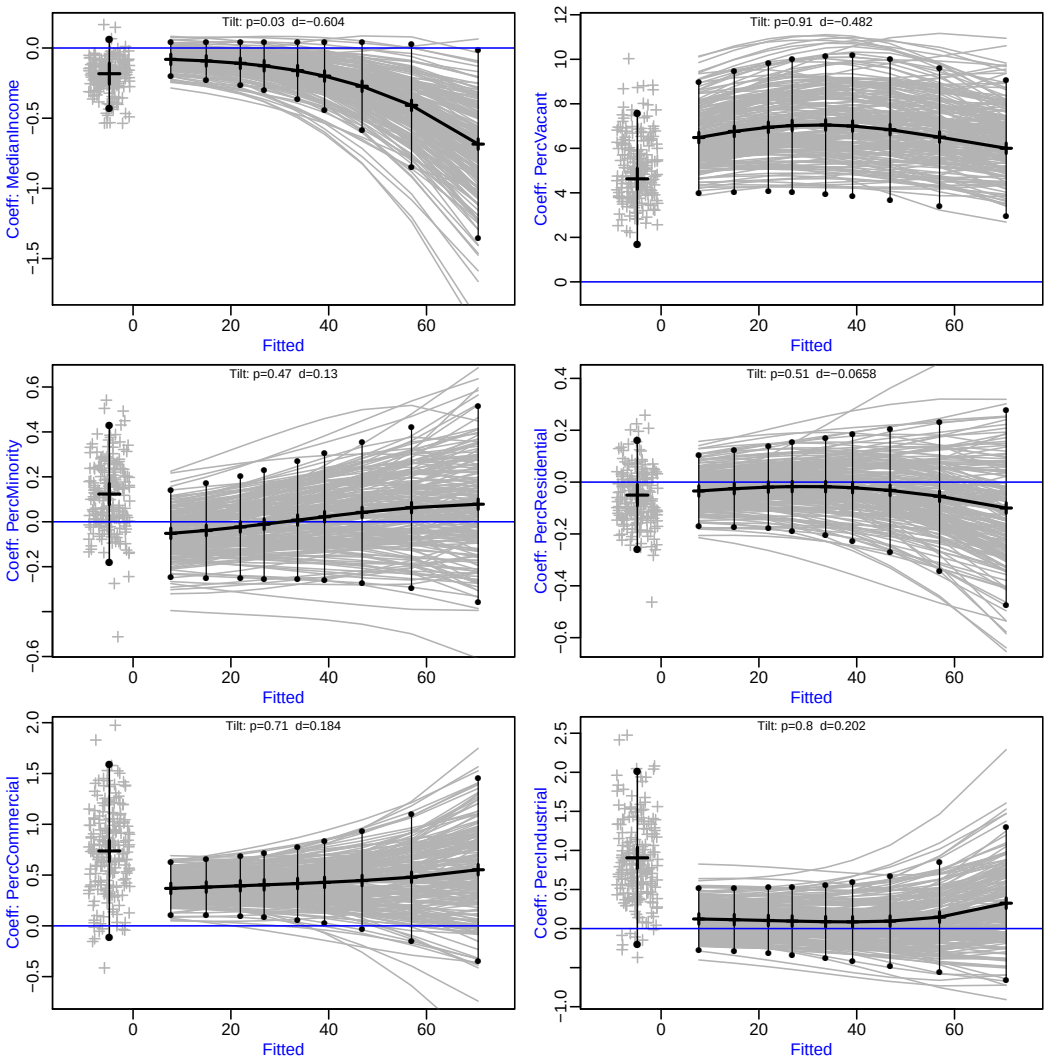


FIG 3. *Misspecification Diagnostics using one focal reweighting variable, the best linear approximation/prediction* `Fitted`*, for all slopes.*

`StreetTotal` of homeless.

- For high values of `PercMinority` $\approx 80\%$, the slope is positive and (weakly) statistically significant: Incrementally more minorities is associated with a higher `StreetTotal` of homeless.

This finding (if real) represents a version of Simpson's paradox: In aggregate, there is no statistically significant association, but, conditional on low and high values of `PercMinority`, there is, and in opposite directions.

In Appendix A.5 we discuss some reasons for the unpredictable behaviors of slopes under reweighting wrt to their own regressors. We also mention a (weak) link to partial additive models with one nonlinear term.

### 5.3 Diagnostics for a Focal Reweighting Variable of Interest (Figure 3)

Next we illustrate a version of the diagnostics that subjects all slopes of a linear regression to a single reweighting variable of interest. The goal is to detect misspecification in any coefficient, and the hope is to do so by reweighting based

on a variable $Z$ that is both powerful and interpretable. Taking a cue from traditional residual diagnostics, we choose the OLS best approximation, $Z = \boldsymbol{\beta}' \vec{\boldsymbol{X}}$. The data version is based on reweighting as function of the OLS estimates of the fitted values, $z_i = \hat{y}_i = \hat{\boldsymbol{\beta}}' \vec{\boldsymbol{x}}_i$.[6] The question is whether any coefficient reveals misspecification when comparing it on data with more low versus more high values of the linear approximation. The expectation is that the gradient of the linear approximation should be a direction of high average response variation and hence may have a higher promise of revealing misspecifications than other directions in regressor space.

Figure 3 shows this diagnostic applied to the LA homeless data, labeling the reweighting variable as `Fitted`. Some observations are as follows:

- The only slope with signs of misspecification is for `MedianIncome` (top left plot), whose tilt test has a p-value of 0.03. This slope achieves mild statistical significance for high values of `Fitted`, which would indicate that the "effect" (if any) of differences in `MedianIncome` matter more for high values of the linear prediction `Fitted`.
- The slope of `PercCommercial` (bottom left plot) shows no signs of misspecification, but it is mildly statistically significant only for low values of `Fitted` due to the lower estimation uncertainty in that range.
- Five of the six plots feature a fan shape of the bootstrap spaghetti bands (exception: `PercVacant`). This indicates that these five slope estimates have greater estimation uncertainty for higher values of `Fitted`.

The last point illustrates that the diagnostic is not only informative about the average level of estimates but also about their estimation uncertainty.

### 5.4 Summary Comments on Reweighting Diagnostics

The reweighting diagnostics proposed here are not meant to replace other types of diagnostics, typically based on residual analysis. They are, however, able to answer questions about quantities of interest and effects of regressors that residual analysis might not. They may also be able to provide insights into the nature of nonlinearities and interactions without explicitly modeling them. Furthermore they are easily augmented with inferential features such as bootstrap spaghetti bands and tests of misspecification with specific interpretations. Finally, they are able to localize regions in regressor space with high or low estimation uncertainty.

## 6. ESTIMATION OF REGRESSION FUNCTIONALS: CANONICAL DECOMPOSITION OF ESTIMATION OFFSETS

We return to the task of building a general framework of plug-in estimation of regression functionals based on iid data. We decompose sampling variability into its two sources, one due to the conditional response distribution, the other due to the randomness of the regressors interacting with misspecification. Along the way we find new characterizations of well-specification of regression functionals.

[6] The estimated slope vector $\hat{\boldsymbol{\beta}}$ is frozen across bootstraps, ignoring a lower-order source of sampling variability.

### 6.1 Regression Data and Plug-In Estimation

We adopt some of the notations and assumptions from Part I, Section 5: Data consist of $N$ iid draws $(Y_i, \vec{\boldsymbol{X}}_i) \sim \boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$; the responses $Y_i$ are collected in a data structure $\boldsymbol{Y} = \{Y_i\}_i$, and the regressors $\vec{\boldsymbol{X}}_i$ in another data structure $\boldsymbol{X} = \{\vec{\boldsymbol{X}}_i\}_i$, called "data frame" in programming languages such as R (2008). We avoid the terms "vector" and "matrix" because in a general theory of regression all variables — responses and regressors — can be of any type and of any dimension.[7] This is why not only $\boldsymbol{X}$ but $\boldsymbol{Y}$ is best thought of as a (random) "data frame." Regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ is any attempt at estimating aspects of the conditional distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. We limit ourselves to regression functionals $\boldsymbol{\theta}(\cdot)$ that allow plug-in estimation $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\widehat{\boldsymbol{P}})$ where $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{P}}_{Y,\vec{\boldsymbol{X}}} = (1/N) \sum \delta_{(Y_i, \vec{\boldsymbol{X}}_i)}$ is the joint empirical distribution. If necessary we may write $\widehat{\boldsymbol{P}}_N$ for $\widehat{\boldsymbol{P}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{N}}$ for $\hat{\boldsymbol{\theta}}$. In addition, we will also need the empirical regressor distribution $\widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}} = (1/N) \sum \delta_{\vec{\boldsymbol{X}}_i}$.

### 6.2 The Conditional Parameter of Model-Trusting Fixed-$X$ Regression

We now define the important notion of a "conditional parameter" for arbitrary regression functionals, thereby providing the target of estimation for fixed-$\boldsymbol{X}$ theories. For OLS slopes this target of estimation is $\boldsymbol{\beta}(\boldsymbol{X}) = \boldsymbol{E}_{\boldsymbol{P}}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$ (Part I, Section 5). We use the idea that fixed-$\boldsymbol{X}$ theories condition on observed regressor observations $\vec{\boldsymbol{X}}_1$, ..., $\vec{\boldsymbol{X}}_N$, collected in the data frame $\boldsymbol{X}$, and define a target of estimation by assuming that the population of $Y$-values at each $\vec{\boldsymbol{X}}_i$ is known: $Y_i|\vec{\boldsymbol{X}}_i \sim \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}_i}$. The joint distribution is then effectively $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$, amounting to partial plug-in of $\widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ for $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ in $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} = \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}$. The conditional parameter for $\boldsymbol{\theta}(\cdot)$ is therefore defined as $\boldsymbol{\theta}(\boldsymbol{X}) = \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}})$. We summarize notations with emphasis on the centerline of the following box:

$$
\boxed{
\begin{array}{lcll}
\boldsymbol{\theta}(\boldsymbol{P}) & = & \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}), & \\
\boldsymbol{\theta}(\boldsymbol{X}) & = & \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}), & \widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}} = (1/N) \sum \delta_{\vec{\boldsymbol{X}}_i}, \\
\hat{\boldsymbol{\theta}} & = & \boldsymbol{\theta}(\widehat{\boldsymbol{P}}). &
\end{array}
}
$$

Note that $\boldsymbol{X}$ and $\widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ contain the same information; the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ is implied and not shown in $\boldsymbol{\theta}(\boldsymbol{X})$. The main points are:

- In model-trusting theories that condition on $\boldsymbol{X}$, the target of estimation is $\boldsymbol{\theta}(\boldsymbol{X})$. They assume $\boldsymbol{\theta}(\boldsymbol{X})$ is the same for all acceptable $\boldsymbol{X}$.
- In model-robust theories that do not condition on $\boldsymbol{X}$, the target of estimation is $\boldsymbol{\theta}(\boldsymbol{P})$, whereas $\boldsymbol{\theta}(\boldsymbol{X})$ is a random quantity (Corollary 6.3 below).

The above definitions can be made more concrete by illustrating them with the specific ways of defining regression functionals of Section 2:

[7] Recall that the typographic difference between $Y$ and $\vec{\boldsymbol{X}}$ is a holdover from Part I, where the response was assumed univariate quantitative.

- Functionals defined through minimization of objective functions:

$$\boxed{\begin{aligned}\boldsymbol{\theta}(\boldsymbol{P}) &= \text{argmin}_{\boldsymbol{\theta}}\ \boldsymbol{E}_{\boldsymbol{P}}[\,\mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})\,],\\ \boldsymbol{\theta}(\boldsymbol{X}) &= \text{argmin}_{\boldsymbol{\theta}}\ \tfrac{1}{N}\textstyle\sum_i \boldsymbol{E}_{\boldsymbol{P}}[\,\mathcal{L}(\boldsymbol{\theta}; Y_i, \vec{\boldsymbol{X}}_i)\,|\,\vec{\boldsymbol{X}}_i],\\ \hat{\boldsymbol{\theta}} &= \text{argmin}_{\boldsymbol{\theta}}\ \tfrac{1}{N}\textstyle\sum_i \mathcal{L}(\boldsymbol{\theta}; Y_i, \vec{\boldsymbol{X}}_i).\end{aligned}}$$

- Functionals defined through estimating equations:

$$\boxed{\begin{aligned}\boldsymbol{\theta}(\boldsymbol{P}):&\quad \boldsymbol{E}_{\boldsymbol{P}}[\,\boldsymbol{\psi}(\boldsymbol{\theta}\, Y, \vec{\boldsymbol{X}})\,] &&= \mathbf{0},\\ \boldsymbol{\theta}(\boldsymbol{X}):&\quad \tfrac{1}{N}\textstyle\sum_i \boldsymbol{E}_{\boldsymbol{P}}[\,\boldsymbol{\psi}(\boldsymbol{\theta}; Y_i, \vec{\boldsymbol{X}}_i)\,|\,\vec{\boldsymbol{X}}_i] &&= \mathbf{0},\\ \hat{\boldsymbol{\theta}}:&\quad \tfrac{1}{N}\textstyle\sum_i \boldsymbol{\psi}(\boldsymbol{\theta}; Y_i, \vec{\boldsymbol{X}}_i) &&= \mathbf{0}.\end{aligned}}$$

These specialize to normal equations for linear OLS by (7).

Summary: Among the three cases in each box, the most impenetrable but also most critical case is the second one. It defines the "conditional parameter" through partial plug-in of the empirical regressor distribution. The conditional parameter is *the target of fixed-$\boldsymbol{X}$ regression for arbitrary types of regression functionals.*

### 6.3 Estimation Offsets

The conditional parameter $\boldsymbol{\theta}(\boldsymbol{X})$ enables us to distinguish between two sources of estimation uncertainty: (1) the conditional response distribution and (2) the marginal regressor distribution. To this end we defined in Part I for linear OLS what we call "estimation offsets." With the availability of $\boldsymbol{\theta}(\boldsymbol{X})$ for regression functionals, these can be defined in full generality:

$$\boxed{\begin{aligned}&\textit{Total EO} &&= \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{P}),\\ &\textit{Noise EO} &&= \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{X}),\\ &\textit{Approximation EO} &&= \boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}(\boldsymbol{P}).\end{aligned}}$$

The total EO is the offset of the plug-in estimate from its population target. The noise EO is the component of the total EO that is due to the conditional distribution $Y|\vec{\boldsymbol{X}}$. The approximation EO is the part due to the randomness of $\vec{\boldsymbol{X}}$ under misspecification. These interpretations will be elaborated in what follows.

**Remark:** We repeat an observation made in Part I, end of Section 5. The approximation EO $\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}(\boldsymbol{P})$ could be misinterpreted as a bias because it is the difference of two targets of estimation. This interpretation is *wrong*. In the presence of misspecification, the approximation EO is a non-vanishing random variable. It will be shown to contribute not a bias to $\hat{\boldsymbol{\theta}}$ but a $N^{-1/2}$ term to the sampling variability of $\hat{\boldsymbol{\theta}}$.

### 6.4 Well-Specification in Terms of Approximation EOs

The approximation EO lends itself for another characterization of well-specification:

**Proposition 6.4:** *Assume $\boldsymbol{P}_{\vec{\boldsymbol{X}}} \mapsto \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}})$ is continuous in the weak topology. Then $\boldsymbol{\theta}(\cdot)$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ iff $\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}(\boldsymbol{P}) = \mathbf{0}$ for all acceptable $\boldsymbol{X}$.*

**Proof**: If $\boldsymbol{\theta}(\cdot)$ is well-specified in the sense of Section 3, then

$$\boldsymbol{\theta}(\boldsymbol{X}) \;=\; \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}) \;=\; \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) \;=\; \boldsymbol{\theta}(\boldsymbol{P}).$$

The converse follows because the empirical regressor distributions $\widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ (for $N \to \infty$) form a weakly dense subset in the set of all regressor distributions, and the regression functional is assumed continuous in this argument. □

A fine point about this proposition is that $\boldsymbol{X}$ is not meant as random but as a variable taking on all acceptable regressor datasets of arbitrarily large sample sizes. On the other hand, here are two consequences when $\boldsymbol{X}$ is random:

**Corollary 6.4:** *Same assumptions as in Proposition 6.4.*

- *Fixed-$\boldsymbol{X}$ and random-$\boldsymbol{X}$ theories estimate the same target iff $\boldsymbol{\theta}(\cdot)$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$.*
- *$\boldsymbol{\theta}(\cdot)$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ iff $\boldsymbol{V}_{\boldsymbol{P}}[\boldsymbol{\theta}(\boldsymbol{X})] = \boldsymbol{0}$ for all acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.*

The first bullet confirms that the notion of well-specification for regression functionals hits exactly the point of agreement between theories that condition on the regressors and those that treat them as random. The second bullet leads the way to the fact that a misspecified regression functional will incur sampling variability originating from the randomness of the regressors.

### 6.5 Deterministic Association Annihilates the Noise EO

While well-specification addresses a vanishing approximation EO, one can also consider the dual concept of a vanishing noise EO. Here is a sufficient condition under which the noise EO vanishes for all regression functionals:

**Proposition 6.5:** *If $Y = f(\vec{\boldsymbol{X}})$ is a deterministic function of $\vec{\boldsymbol{X}}$, then $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{X}) = \boldsymbol{0}$ for all regression functionals.*

**Proof**: The conditional response distribution is $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}} = \delta_{y=f(\vec{\boldsymbol{x}})}$, hence the joint distribution formed from $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}$ and $\widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ is $\widehat{\boldsymbol{P}}$: $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}} = \widehat{\boldsymbol{P}}$. It follows that $\boldsymbol{\theta}(\boldsymbol{X}) = \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \widehat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}(\widehat{\boldsymbol{P}}) = \hat{\boldsymbol{\theta}}$. □

The proposition illustrates the fact that the noise EO is due to "noise", that is, variability of $Y$ conditional on $\vec{\boldsymbol{X}}$. Thus, although less transparent than in linear OLS, the conditional response distribution $Y|\vec{\boldsymbol{X}}$ is the driver of the noise EO.

### 6.6 Well-Specification and Influence Functions

This section introduces influence functions for regression functionals which will prove useful for approximations in Section 6.7 and for asymptotic decompositions in Section 7. For background on influence functions see, for example, Hampel et al. (1986) and Rieder (1994).

The influence function is a form of derivative on the space of probability distributions, which makes it an intuitive tool to characterize well-specification of regression functionals: If $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}})$ is constant in the argument $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ at a fixed $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, then this means intuitively that the "partial derivative" wrt $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ vanishes.

The definition of the full influence function of $\boldsymbol{\theta}(\cdot)$ is as follows:

$$\boldsymbol{IF}(y,\vec{\boldsymbol{x}}) \;=\; \frac{d}{dt}\Big|_{t=0} \boldsymbol{\theta}\left((1-t)\boldsymbol{P} + t\delta_{(y,\vec{\boldsymbol{x}})}\right). \tag{10}$$

We omit $\boldsymbol{\theta}(\cdot)$ as well as $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ as arguments of $\boldsymbol{IF}(y,\vec{\boldsymbol{x}})$ because both will be clear from the context, except for one occasion in Appendix A.3 where we write $\boldsymbol{IF}(y,\vec{\boldsymbol{x}};\boldsymbol{P})$. More relevant is the following definition of the partial influence function of $\boldsymbol{\theta}(\cdot)$ with regard to the regressor distribution:

$$\boldsymbol{IF}(\vec{\boldsymbol{x}}) \;=\; \frac{d}{dt}\Big|_{t=0} \boldsymbol{\theta}\big(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes ((1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\delta_{\vec{\boldsymbol{x}}})\big). \tag{11}$$

For derivations of the following Lemma and Proposition, see Appendix A.3.

**Lemma 6.6:** $\boldsymbol{IF}(\vec{\boldsymbol{x}}) \;=\; \boldsymbol{E}_{\boldsymbol{P}}\big[\,\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}\,\big]$.

**Proposition 6.6:** *A regression functional $\boldsymbol{\theta}(\cdot)$ with an influence function at $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ iff* $\boldsymbol{IF}(\vec{\boldsymbol{x}}) = \boldsymbol{0} \;\; \forall \vec{\boldsymbol{x}}$.

### 6.7 Approximating Estimation Offsets with Influence Functions

For linear OLS, Definition and Lemma 5 in Part I exhibited an intuitive correspondence between the total, noise and approximation EO on the one hand and the population residual, the noise and the nonlinearity on the other hand. No such direct correspondence exists for general types of regression. The closest general statement about EOs is in terms of approximations based on influence functions. Assuming asymptotic linearity of $\boldsymbol{\theta}(\cdot)$, the EOs have the following approximations to order $o_P(N^{-1/2})$:

$$\boxed{\begin{array}{lll} \textit{Total EO:} & \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}(\boldsymbol{P}) & \approx \frac{1}{N}\sum_i \boldsymbol{IF}(Y_i,\vec{\boldsymbol{X}}_i), \\ \textit{Noise EO:} & \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}(\boldsymbol{X}) & \approx \frac{1}{N}\sum_i \Big(\boldsymbol{IF}(Y_i,\vec{\boldsymbol{X}}_i) - \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}}_i)|\vec{\boldsymbol{X}}_i]\Big), \\ \textit{Approx. EO:} & \boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}(\boldsymbol{P}) & \approx \frac{1}{N}\sum_i \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}}_i)|\vec{\boldsymbol{X}}_i]. \end{array}} \tag{12}$$

These approximations (12) lead straight to the CLTs of the next section.

## 7. MODEL-ROBUST CENTRAL LIMIT THEOREMS DECOMPOSED

### 7.1 CLT Decompositions Based on Influence Functions

As in Section 6.7 assume the regression functional $\boldsymbol{\theta}(\boldsymbol{P})$ is asymptotically linear with influence function $\boldsymbol{IF}(y,\vec{\boldsymbol{x}})$ and partial influence function $\boldsymbol{IF}(\vec{\boldsymbol{x}}) = \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})\,|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}]$. The EOs obey the following CLTs:

$$\boxed{\begin{array}{lll} \sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{P})) & \xrightarrow{\mathcal{D}} & \mathcal{N}\Big(\boldsymbol{0},\, \boldsymbol{V}_{\boldsymbol{P}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})]\Big), \\ \sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{X})) & \xrightarrow{\mathcal{D}} & \mathcal{N}\Big(\boldsymbol{0},\, \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{V}_{\boldsymbol{P}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]]\Big), \\ \sqrt{N}\,(\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}(\boldsymbol{P})) & \xrightarrow{\mathcal{D}} & \mathcal{N}\Big(\boldsymbol{0},\, \boldsymbol{V}_{\boldsymbol{P}}[\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]]\Big). \end{array}}$$

These are immediate consequences of the assumed asymptotic linearities. The asymptotic variances of the EOs follow the canonical decomposition

$$\boldsymbol{V_P}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})] \;=\; \boldsymbol{E_P}[\boldsymbol{V_P}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,] + \boldsymbol{V_P}[\boldsymbol{E_P}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,],$$

the three terms being the asymptotic variance-covariance matrices of the total, the noise and the approximation EO, respectively. Implicit in this Pythagorean formula is that $\boldsymbol{IF}(Y,\vec{\boldsymbol{X}}) - \boldsymbol{E_P}[\boldsymbol{IF}(Y,|\vec{\boldsymbol{X}})$ and $\boldsymbol{E_P}[\boldsymbol{IF}(Y,|\vec{\boldsymbol{X}})$ are orthogonal to each other, which implies by (12) that the noise EO and the approximation EO are asymptotically orthogonal. Asymptotic orthogonalities based on conditioning are well-known in semi-parametric theory. For linear OLS this orthogonality holds exactly for finite $N$ due to (6) and (13) in Part I: $\boldsymbol{V_P}[\hat{\boldsymbol{\beta}}\text{–}\boldsymbol{\beta}(\boldsymbol{X}), \boldsymbol{\beta}(\boldsymbol{X})\text{–}\boldsymbol{\beta}(\boldsymbol{P})] = \boldsymbol{0}$.

The following corollary is a restatement of Proposition 6.6, but enlightened by the fact that it relies on the asymptotic variance of the approximation EO.

**Corollary 7.1:** *The regression functional $\boldsymbol{\theta}(\cdot)$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ iff the asymptotic variance of the approximation EO vanishes for all acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.*

**Proof**: Using careful notation the condition says $\boldsymbol{V}_{\boldsymbol{P}_{\vec{\boldsymbol{X}}}}[\boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})|\vec{\boldsymbol{X}}]] = \boldsymbol{0}$ for all acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. This in turn means $\boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \boldsymbol{0}$ for all $\vec{\boldsymbol{x}}$, which is the condition of Proposition 6.6. □

### 7.2 CLT Decompositions for EE Functionals

For EE functionals the influence function is $\boldsymbol{IF}(y,\vec{\boldsymbol{x}}) = \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\boldsymbol{\psi}(\boldsymbol{\theta};y,\vec{\boldsymbol{x}})$ where $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{P})$ and $\boldsymbol{\Lambda}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{E_P}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})]$ is the Jacobian of size $q \times q$, $q = \dim(\boldsymbol{\psi}) = \dim(\boldsymbol{\theta})$. Then the CLTs specialize to the following:

$$\begin{aligned}
\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \;&\xrightarrow{\mathcal{D}}\; \mathcal{N}\Big(\boldsymbol{0},\, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{V_P}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})]\;\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}\Big) \\
\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{X})) \;&\xrightarrow{\mathcal{D}}\; \mathcal{N}\Big(\boldsymbol{0},\, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{E_P}[\boldsymbol{V_P}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,]\;\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}\Big) \\
\sqrt{N}\,(\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}) \;&\xrightarrow{\mathcal{D}}\; \mathcal{N}\Big(\boldsymbol{0},\, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{V_P}[\boldsymbol{E_P}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,]\;\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}\Big)
\end{aligned}$$

The first line is Huber's (1967, Section 3) result. The asymptotic variances have the characteristic sandwich form. It is natural that they are related according to

$$\boldsymbol{V_P}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})] \;=\; \boldsymbol{E_P}[\boldsymbol{V_P}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,] + \boldsymbol{V_P}[\boldsymbol{E_P}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,],$$

where on the right side the first term relates to the noise EO and the second term to the approximation EO.

Linear OLS is a special case with $\boldsymbol{\psi}(\boldsymbol{\beta};y,\vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}}\vec{\boldsymbol{x}}'\boldsymbol{\beta} - \vec{\boldsymbol{x}}y$, $\boldsymbol{\Lambda} = \boldsymbol{E_P}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']$, $\boldsymbol{IF}(y,\vec{\boldsymbol{x}}) = \boldsymbol{E_P}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}(\vec{\boldsymbol{x}}\vec{\boldsymbol{x}}'\boldsymbol{\beta} - \vec{\boldsymbol{x}}y)$, and hence the CLTs of Part I, Proposition 7.

### 7.3 Implications of the CLT Decompositions

We address once again potential confusions relating to different notions of bias. Misspecification, in traditional parametric modeling, is sometimes called "model bias" which, due to unfortunate terminology, may suggest a connection to estimation bias, $\boldsymbol{E_P}[\hat{\boldsymbol{\theta}}_N] - \boldsymbol{\theta}(\boldsymbol{P})$. Importantly, there is no connection between the

two notions of bias. Estimation bias typically vanishes at a rate faster than $N^{-1/2}$ and does not contribute to standard errors derived from asymptotic variances. Model bias, on the other hand, which is misspecification, generates in conjunction with the randomness of the regressors a contribution to the standard error, and this contribution is asymptotically of order $N^{-1/2}$, the same order as the better known contribution due to the conditional noise in the response. This is what the CLT decomposition shows. It also shows that the two sources of sampling variability are asymptotically orthogonal. — In summary:

*Model bias/misspecification does not create estimation bias; it creates sampling variability to the same order as the conditional noise in the response.*

## 8. PLUG-IN/SANDWICH ESTIMATORS VERSUS $M$-OF-$N$ BOOTSTRAP ESTIMATORS OF STANDARD ERROR

### 8.1 Plug-In Estimators are Limits of $M$-of-$N$ Bootstrap Estimators

In Part I, Section 8, it was indicated that for linear OLS there exists a connection between two ways of estimating asymptotic variance: the sandwich estimator for sample size $N$ is the limit of the $M$-of-$N$ bootstrap as $M \to \infty$, where bootstrap is the kind that resamples $x$-$y$ cases rather than residuals. This connection holds at a general level: all plug-in estimators of standard error are limits of bootstrap in this sense.

The crucial observation of Part I goes through as follows: The $M$-of-$N$ bootstrap is iid sampling of $M$ observations from some distribution, hence there must hold a CLT as the resample size grows, $M \to \infty$. The distribution being (re)sampled is the empirical distribution $\widehat{\boldsymbol{P}}_N = (1/N)\sum \delta_{(y_i,\vec{\boldsymbol{x}}_i)}$, where $N$ is fixed but $M \to \infty$.[8] Therefore, the following holds for bootstrap resampling of any well-behaved statistical functional, be it in a regression context or not:

**Proposition 8.1:** *Assume the regression functional $\boldsymbol{\theta}(\cdot)$ is asymptotically normal for a sufficiently rich class of joint distributions $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ with acceptable regressor distributions $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ as follows:*

$$N^{1/2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{N}} - \boldsymbol{\theta}(\boldsymbol{P})) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(\mathbf{0},\ \boldsymbol{AV}[\boldsymbol{P};\boldsymbol{\theta}(\cdot)]\right) \qquad (N \to \infty).$$

*Let a fixed dataset of size $N$ with acceptable regressors be represented by the empirical measure $\widehat{\boldsymbol{P}}_N$. Then a CLT holds for the $M$-of-$N$ bootstrap as $M \to \infty$, with an asymptotic variance obtained by plug-in. Letting $\boldsymbol{\theta}^*_M = \boldsymbol{\theta}(\boldsymbol{P}^*_{\boldsymbol{M}})$ where $\boldsymbol{P}^*_{\boldsymbol{M}}$ is the empirical distribution of a resample of size $M$ from $\widehat{\boldsymbol{P}}_N$, we have:*

$$M^{1/2}\left(\boldsymbol{\theta}^*_M - \hat{\boldsymbol{\theta}}_{\boldsymbol{N}}\right) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(\mathbf{0},\ \boldsymbol{AV}[\widehat{\boldsymbol{P}}_N;\boldsymbol{\theta}(\cdot)]\right) \qquad (M \to \infty, \quad N \text{ fixed}).$$

The proposition contains its own proof. The following is the specialization to EE functionals where the asymptotic variance has the sandwich form:

**Corollary 8.1:** *The plug-in sandwich estimator for an EE functional is the asymptotic variance estimated by the $M$-of-$N$ bootstrap in the limit $M \to \infty$ for a fixed sample of size $N$.*

[8] This causes ever more ties at $M$ grows.

### 8.2 Arguments in Favor of $M$-of-$N$ Bootstrap Over Plug-In Estimators

A natural next question is whether the plug-in/sandwich estimator is to be preferred over $M$-of-$N$ bootstrap estimators, or whether there is a reason to prefer some form of $M$-of-$N$ bootstrap. In the latter case the follow-up question would be how to choose the resample size $M$. While we do not have any recommendations for choosing a specific $M$, there exist various arguments in favor of some $M$-of-$N$ bootstrap over plug-in/sandwich estimation of standard error.

A first argument is that bootstrap is more flexible in that it lends itself to various forms of confidence interval construction that grant higher order accuracy of coverage. See, for example, Efron and Tibshirani (1994) and Hall (1992).

A second argument is related to the first but in a different direction: Bootstrap can be used to diagnose whether the sampling distribution of a particular functional $\boldsymbol{\theta}(\cdot)$ is anywhere near asymptotic normality for a given sample size $N$. This can be done by applying normality tests to simulated bootstrap values $\boldsymbol{\theta}_b^*$ $(b = 1, ..., B)$, or by displaying these values in a normal quantile plot.

A third argument is that there exists theory that shows bootstrap to work for very small $M$ compared to $N$ in some situations where even conventional $N$-of-$N$ bootstrap does not work. (See Bickel, Götze and van Zwet (1997) following Politis and Romano (1994) on subsampling.) It seems therefore unlikely that the limit $M \to \infty$ for fixed $N$ will yield any form of superiority to bootstrap with finite $M$.

A fourth argument is based on a result by Buja and Stuetzle (2016) which will be elaborated in the remainder of this section. The result states that so-called "$M$-bagged functionals" have low complexity in a certain sense, the lower the smaller the resample size $M$. In this sense, the limit $M \to \infty$ is the most complex choice that is more likely to behave erratically than all choices of finite $M$. We next define "bagged functionals", explain the meaning of "complexity", and describe the link of this result to the issue of "bootstrap versus plug-in/sandwich estimators".

"Bagging" is "bootstrap aggregating" as introduced by Breiman (1996). While bootstrap and bagging are applied in practice to empirical distributions $\widehat{\boldsymbol{P}}_N$, we need a definition of bagging for arbitrary distributions. Following Buja and Stuetzle (2016) the $M$-**bagged** functional $\boldsymbol{\theta}_M^B(\cdot)$ associated with $\boldsymbol{\theta}(\cdot)$ is

$$\boldsymbol{\theta}_M^B(\boldsymbol{P}) = \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\theta}(\boldsymbol{P}_M)], \qquad \boldsymbol{P}_M = \frac{1}{M}\sum_{i=1}^{M} \delta_{(Y_i, \vec{\boldsymbol{X}}_i)}, \quad (Y_i, \vec{\boldsymbol{X}}_i) \sim \boldsymbol{P} \text{ iid.}$$

The $M$-bagged functional $\boldsymbol{\theta}_M^B(\boldsymbol{P})$ is the average of the functional $\boldsymbol{\theta}(\cdot)$ over empirical measures of $M$ iid observations from $\boldsymbol{P}$. These $M$ "observations" are not actual observations but rather hypothetical iid draws from $\boldsymbol{P}$ for the purposes of a mathematical construction. The definition of an $M$-bagged statistical functional provides a target of estimation when the bag size $M$ is held fixed and $N \to \infty$.

Bagged functionals have the remarkable property that their von Mises expansions are finite, irrespective of the nature of the original functional $\boldsymbol{\theta}(\cdot)$. Von Mises expansions are based on generalized influence functions of higher order, where the first order term is the average of the first order influence function (Section 6.7) that drives asymptotic normality (Section 7). Higher order terms reflect "interactions" similar to U-statistics (actually, V-statistics) based on pairs, triples, quadruples, ... of observations. It is intuitively clear that an $M$-bagged functional cannot have interactions of order higher than $M$. This was confirmed by Buja and Stuetzle (2016) to whom we refer for the explicit form of the expansion:

**Proposition:** *$M$-bagged functionals have von Mises expansions of length $M+1$.*

This proposition suggests a notion of complexity for bagged functionals that is analogous to the natural notion of complexity of polynomials: A polynomial of higher order is higher in complexity, and, similarly, a bagged functional of higher bag size $M$ is of higher complexity. In both cases it is the length of the expansion that is a measure of "complexity."

The connection of $M$-bagged functionals to the present issue is that bootstrap estimators of sampling variance can be seen as plug-in estimators of certain variance functionals. To see this we define the normalized $M$-bootstrap variance functional $\boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)]$ associated with $\boldsymbol{\theta}(\cdot)$, which is obtained essentially by replacing expectation with variance in the definition of the $M$-bagged functional $\boldsymbol{\theta}_M^B(\boldsymbol{P})$:

$$\boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)] \;=\; M\cdot\boldsymbol{V}_{\boldsymbol{P}}[\boldsymbol{\theta}(\boldsymbol{P}_M)].$$

The connection with asymptotic variance is simply

$$\boldsymbol{AV}[\boldsymbol{P};\boldsymbol{\theta}(\cdot)] \;=\; \lim_{M\to\infty}\boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)].$$

Both the $M$-of-$N$ bootstrap estimator and the plug-in/sandwich estimator of standard error are obtained by replacing $\boldsymbol{P}$ with the empirical distribution $\widehat{\boldsymbol{P}}_N$ of an observed sample of size $N$, and renormalizing accordingly. For scalar functionals $\boldsymbol{\theta}(\cdot)$ the standard error estimates are therefore:

$$\hat{\boldsymbol{SE}}^{boot}_{M\text{-of-}N} = \frac{1}{\sqrt{N}}\boldsymbol{BV}_M[\widehat{\boldsymbol{P}}_N;\boldsymbol{\theta}(\cdot)]^{1/2}, \qquad \hat{\boldsymbol{SE}}^{sand}_{N} = \frac{1}{\sqrt{N}}\boldsymbol{AV}[\widehat{\boldsymbol{P}}_N;\boldsymbol{\theta}(\cdot)]^{1/2}.$$

In this interpretation both the $M$-of-$N$ bootstrap and the plug-in/sandwich estimator are plug-in estimators, the difference being in the variance functionals to which plug-in is applied. These variance functionals are, respectively:

$$\boldsymbol{P}\mapsto\boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)] \;\text{ and }\; \boldsymbol{P}\mapsto\boldsymbol{AV}[\boldsymbol{P};\boldsymbol{\theta}(\cdot)]. \tag{13}$$

The bootstrap variance functionals are trivially shown to be a "function of bagged functionals" by rewriting them as

$$\boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)] \;=\; M\cdot\left(\,\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\theta}(\boldsymbol{P}_M)^2] - \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\theta}(\boldsymbol{P}_M)]^2\,\right),$$

which are a function of the $M$-bagged versions of the functionals $\boldsymbol{\theta}(\cdot)^2$ and $\boldsymbol{\theta}(\cdot)$:

$$(\boldsymbol{\theta}^2)^B_M(\boldsymbol{P}) = \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\theta}(\boldsymbol{P}_M)^2] \quad\text{and}\quad \boldsymbol{\theta}^B_M(\boldsymbol{P}) = \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\theta}(\boldsymbol{P}_M)].$$

Therefore, if the complexity of $M$-bagged functionals is measured by the length of the von Mises expansion, the $M$-bootstrap variance functional $\boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)]$ is the lower in complexity the smaller the resampling size $M$ is.

The practical consequence of the preceding constructions is as follows: Lesser complexity of a standard error estimate tends to imply greater sampling stability, hence bootstrap standard errors are likely to be more stable than plug-in/sandwich estimators of standard error. The greatest stability gains are obtained for small resampling sizes compared to $N$, if the results of Politis and Romano (1994) for sub-sampling and Bickel et al. (1997) for bootstrap sampling are a guide. However, the latter authors also show that it is the conventional $N$-of-$N$ bootstrap that is semi-parametrically efficient for the intended standard error of a sample of size $N$. In summary, bootstrap standard errors of any resample size should be preferable to plug-in/sandwich standard errors.

## 9. SUMMARY AND CONCLUSION

This article completes important aspects of the program set out in Part I. It pursues the idea of model robustness to its conclusion for arbitrary types of regression based on iid observations. The notion of model robustness coalesces into a model-free theory where all quantities of interest are statistical functionals, called "regression functionals", and models take on the role of heuristics to suggest objective functions whose minima define regression functionals defined on largely arbitrary joint $(Y, \vec{\boldsymbol{X}})$ distributions. In this final section we recount the path that makes the definition of well-specification for regression functionals compelling.

To start, an important task of the present article has been to extend the two main findings of Part I from linear OLS to arbitrary types of regression. The findings are that nonlinearity and randomness of the regressors interact ("conspire")

(1) to cause the target of estimation to depend on the regressor distribution;
(2) to cause $N^{-1/2}$ sampling variability to arise that is wholly different from the sampling variability caused by the conditional noise in the response.

It was intuitively clear that these effects would somehow carry over from linear OLS to all types of regression, but it wasn't clear what would take the place of "nonlinearity," a notion of first order misspecification peculiar to fitting linear equations and estimating linear slopes. In attempting to generalize Part I, a vexing issue is that one is looking for a framework free of specifics of fitted equations and additive stochastic components of the response. Attempts at directly generalizing the notions of "nonlinearity" and "noise" of Part I lead to dead ends of unsatisfactory extensions that are barely more general than linear OLS. This raises the question to a level of generality in which there is very little air to breathe: the objects that remain are a regression functional $\boldsymbol{\theta}(\cdot)$ and a joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$. Given these two objects, what do mis- and well-specification mean? An answer, maybe *the* answer, is arrived at by casting regression in the most fundamental way possible: *Regression is the attempt to describe the conditional response distribution* $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. This interpretation sweeps away idiosyncratic structure of special cases. It also suggests taking the joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ apart and analyzing the issue of mis- and well-specification in terms of $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, as well as $\boldsymbol{\theta}(\cdot)$, the quantities of interest. The solution, finally, to

- establishing a compelling notion of mis- and well-specification at this level of generality, and
- extending (1) and (2) above to arbitrary types of regression,

is to look no further and use the "conspiracy effect" (1) as the definition: Misspecification means dependence of the regression functional on the regressor distribution. Conversely, well-specification means the regression functional does not depend on the regressor distribution; it is a property of the conditional response distribution alone.

The "conspiracy effect" (2) above is now a corollary of the definition: If the functional is not constant across regressor distributions, it will incur random variability on empirical regressor distributions, and this at the familiar rate $N^{-1/2}$.

The link between the proposed definition and conventional ideas of mis-/well-specification is as follows: Because most regressions consist of fitting some functional form of the regressors to the response, misspecification of the functional

form is equivalent to misspecification of its parameters viewed as regression functionals: depending on where the regressors fall, the misspecified functional form needs adjustment of its parameters to achieve the best approximation over the distribution of the regressors.

Well-specification being an ideal, in reality we always face degrees of misspecification. Acknowledging the universality of misspecification, however, does *not* justify carelessness in practice. It is mandatory to perform diagnostics and, in fact, we proposed a type of diagnostic in Sections 4 and 5 tailored to the present notion of mis-/well-specification. The diagnostic consists of checking the dependence of regression functionals on the regressor distribution by systematically perturbing the latter, not by shifting or otherwise moving it, but by reweighting it. Reweighting has the considerable advantage over other forms of perturbation that it applies to all variable types, not just quantitative ones.

While the reality of misspecification imposes a duty to perform diagnostics, there is also an argument to be made to feel less guilty about choosing simpler models over more complex ones. One reason is that the reweighting diagnostic permits localization of models and thereby enables a systematic exploration of local best approximations, always in terms of model parameters interpreted as regression functionals. As shown in Sections 5.1-5.3, this possibility vastly extends the expressive power of models beyond that of a single model fit.

Finally, there is an argument to be made in favor of using statistical inference that is model-robust, and to this end one can use $x$-$y$ bootstrap estimators or plug-in/sandwich estimators of standard errors. Between the two, one can give arguments in favor of bootstrap over plug-in/sandwich estimators. Most importantly, though, both approaches to inference are in accord with the insight that misspecification forces us to treat regressors as random.

**Acknowledgments**: We are grateful to Gemma Moran and Bikram Karmakar for their help in the generalizations of Section 6.2, and to Hannes Leeb for pointing out the source of the Tukey quote shown before the introduction.

## REFERENCES

[1] Basu, A., Harris, I. R., Hjort, N. L.., Jones, M. C. (1998). Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika* **85** (3), 549-559.

[2] Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid Post-Selection Inference. *The Annals of Statistics* **41** (2), 802–837.

[3] Berk, R. H. and Kriegler, B. and Yilvisaker, D. (2008). Counting the Homeless in Los Angeles County. in *Probability and Statistics: Essays in Honor of David A. Freedman*, Monograph Series for the Institute of Mathematical Statistics, D. Nolan and S. Speed (eds.)

[4] Bickel, P. J. and Götze, F. and van Zwet, W. R. (1997). Resampling Fewer than $n$ Observations: Gains, Losses, and Remedies for Losses. *Statistica Sinica* **7**, 1–31.

[5] Breiman, L. (1996). Bagging Predictors. *Machine Learning* **24**, 123-140.

[6] Buja, A. and Stuetzle, W. (2016, 2001). Smoothing Effects of Bagging: Von Mises Expansions of Bagged Statistical Functionals. *arXiv:1612.02528*

[7] Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press.

[8] Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction and Estimation. *Journal of the American Statistical Association* **102** (477), 359–378.

[9] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* (Springer Series in Statistics) New York, NY: Springer-Verlag.

[10] Hampel, F. R. and Ronchetti, E. M. and Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach based on Influence Functions.* New York, NY: Wiley.

[11] Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **50** (4), 1029–1054.

[12] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, London: Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

[13] Harrison, X. and Rubinfeld, X. (1978). Hedonic Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management* **5**, 81–102.

[14] Hartigan J. A. (1983). Asymptotic Normality of Posterior Distributions. In: *Bayes Theory*, Chapter 11, Springer Series in Statistics. New York, NY: Springer.

[15] Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica* **46** (6), 1251-1271.

[16] Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* **35** (1) 73–101.

[17] Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimation under Nonstandard Conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Berkeley: University of California Press, 221–233.

[18] Imbens, G. W. and Rubing, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York, NY: Cambridge University Press.

[19] Kent, J. (1982). Robust Properties of Likelihood Ratio Tests. *Biometrika* **69** (1), 19–27.

[20] Kuchibhotla, A. K and Brown, L. D. and Buja, A. (2018). Model-free Study of Ordinary Least Squares Linear Regression. *arXiv e-prints:1809.10538*.

[21] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73** (1), 13-22.

[22] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference (2nd ed.)*. New York, NY: Cambridge University Press.

[23] Peters, J. and Bühlmann, P. and Meinshausen, N. (2016). Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals *Journal of the Royal Statistical Society, B* **78**, Part 5, 947–1012.

[24] Politis, D. N. and Romano, J. P. (1994). A General Theory for Large Sample Confidence Regions based on Subsamples under Minimal Assumptions. *The Annals of Statistics* **22**, 2031–2050.

[25] R Development Core Team (2008). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing,* Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[26] Rieder, H. (1994). *Robust Asymptotic Statistics*, New York, NY: Springer-Verlag.

[27] Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics* **33** (1), 1-67.

[28] White, H. (1980a). Using Least Squares to Approximate Unknown Regression Functions. *International Economic Review* **21** (1), 149-–170.

[29] White, H. (1980b). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **48**, 817-–838.

[30] White, H. (1981). Consequences and Detection of Misspecified Nonlinear Regression Models. *Journal of the American Statistical Association* **76** (374), 419-–433.

[31] White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**, 1–25.

[32] White, H. (1994). *Estimation, Inference and Specification Analysis.* Econometric Society Monographs No. 22. Cambridge, GB: Cambridge University Press.

# APPENDIX

## A.1 Assumptions

When defining a regression functional $\boldsymbol{\theta}(\boldsymbol{P})$, one needs to specify a set $\mathcal{P}$ of joint distributions $\boldsymbol{P}=\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ for which the functional is defined. This set can be specific to the functional in several ways. Here is a list of conditions on $\mathcal{P}$ that will be assumed as needed:

a) Expectations $\boldsymbol{E}_{\boldsymbol{P}}[f(\vec{\boldsymbol{X}},Y)]$ exist as needed for all distributions in $\boldsymbol{P}\in\mathcal{P}$.
b) If the regression functional derives from parameters of fitted equations, it will be assumed that the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ grants identifiabiliy of the fitted parameters, as when strict collinearity of the regressor distribution needs to be excluded in order to uniquely fit linear equations. If this is the case we will say $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is an "*acceptable*" regressor distribution.
c) With b) in mind, we will assume that if a regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is acceptable, a mixture $\alpha\boldsymbol{P}_{\vec{\boldsymbol{X}}}+(1-\alpha)\boldsymbol{P}'_{\vec{\boldsymbol{X}}}$ with any other distribution $\boldsymbol{P}'_{\vec{\boldsymbol{X}}}$, subject to a) above, will also be acceptable, the reason being that mixing can only enlarge but not diminish the support of the distribution, hence identifiability will be inherited from $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ irrespective of $\boldsymbol{P}'_{\vec{\boldsymbol{X}}}$.
d) Item c) ensures that the set of acceptable regressor distributions is so rich that $E_{\boldsymbol{P}_{\vec{\boldsymbol{X}}}}[f(\vec{\boldsymbol{X}}]=0$ for all acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ entails $f\equiv 0$. Reason: Mix with atoms at arbitrary locations.
e) For working models $\{\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}\colon\ \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ (not treated as correct) it will be assumed $\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}\otimes\boldsymbol{P}_{\vec{\boldsymbol{X}}}\in\mathcal{P}$ for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and all acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.
f) Where conditional model densities $q(y|\vec{\boldsymbol{x}};\boldsymbol{\theta})$ of the response appear, they will be densities with regard to some dominating measure $\nu(dy)$.
g) For plug-in estimation it will be required that for $N$ iid draws $(Y_i,\vec{\boldsymbol{X}}_i)\sim\boldsymbol{P}\in\mathcal{P}$ the empirical distribution

$$\widehat{\boldsymbol{P}}=\widehat{\boldsymbol{P}}_N=(1/N)\sum\delta_{(Y_i,\vec{\boldsymbol{X}}_i)}$$

is in $\mathcal{P}$ with limiting probability 1 as $N\to\infty$. For example, one needs $N\geq p+1$ non-collinear observations in order to fit a linear equation with intercept.
h) To form influence functions for regression functionals, it will be assumed that for $\boldsymbol{P}\in\mathcal{P}$ and $(y,\vec{\boldsymbol{x}})\in\mathcal{X}\times\mathcal{Y}$ we have $(1-t)\boldsymbol{P}+t\delta_{y,\vec{\boldsymbol{x}}}\in\mathcal{P}$ for $0<t<1$.
i) $\mathcal{P}$ will be assumed to be convex, hence closed under finite mixtures.

## A.2 Proper Scoring Rules, Bregman Divergences and Entropies

*A.2.1 General Theory:* We describe objective functions called "proper scoring rules", which generalize negative log-likelihoods, and associated Bregman divergences, which generalize Kullback-Leibler (K-L) divergences. Proper scoring rules can be used to extend the universe of regression functionals based on working models. For insightful background on proper scoring rules, see Gneiting and Raftery (2007) (but note two reversals of conventions: theirs are gain functions where ours are loss functions, and their order of arguments is switched from ours).

We begin by devising discrepancy measures between pairs of distributions based on axiomatic requirements that can be gleaned from two properties of K-L divergences: Fisher consistency at the working model and availability of plug-in for "empirical risk minimization" (machine learning terminology). The resulting functionals will be called "proper scoring functionals."

Denoting a discrepancy measure between two distributions by $D(\boldsymbol{P}, \boldsymbol{Q})$, the intended roles of the two arguments are that $\boldsymbol{P}$ is the actual data distribution and $\boldsymbol{Q}$ is a member of a working model. Fisher consistency of a minimum discrepancy functional follows from the following requirements:

(A) $D(\boldsymbol{P}, \boldsymbol{Q}) \geq 0$ with equality iff $\boldsymbol{P} = \boldsymbol{Q}$.

The "only if" part in the second clause is essential. Other properties such as symmetry and triangle inequalities are not needed and would be too restrictive.

As for the availability of plug-in estimation, it would follow from a structural property such as dependence of $D(\boldsymbol{P}, \boldsymbol{Q})$ on $\boldsymbol{P}$ only through its expectation $\boldsymbol{E}_{\boldsymbol{P}}[...]$, whose plug-in estimate is the empirical mean. Other types of plug-in exist, for example for quantiles, in particular medians. Yet other cases, such as Hellinger distances, require density estimation for plug-in, which adds a layer of complexity. In what follows we will impose the strong condition that $D(\boldsymbol{P}, \boldsymbol{Q})$ depends on $\boldsymbol{P}$ essentially only through $\boldsymbol{E}_{\boldsymbol{P}}[\cdot]$, but this requirement only concerns the part of $D(\boldsymbol{P}, \boldsymbol{Q})$ that is relevant for minimization over working model distributions $\boldsymbol{Q}$. We can use the K-L divergence as a guide: In

$$D_{KL}(\boldsymbol{P}, \boldsymbol{Q}) \;=\; \boldsymbol{E}_{\boldsymbol{P}}[-\log \frac{q(Y)}{p(Y)}] \;=\; \boldsymbol{E}_{\boldsymbol{P}}[-\log q(Y)] - \boldsymbol{E}_{\boldsymbol{P}}[-\log p(Y)],$$

the second term requires for plug-in a density estimate of $p(y)$, but this term does not depend on $\boldsymbol{Q}$, hence is irrelevant for minimization over $\boldsymbol{Q}$. By analogy we impose the following structural form on the discrepancy measure:

(B) $D(\boldsymbol{P}, \boldsymbol{Q}) \;=\; \boldsymbol{E}_{\boldsymbol{P}}[S(Y, \boldsymbol{Q})] - H(\boldsymbol{P})$ .

This condition, combined with condition (A), constrains $D(\boldsymbol{P}, \boldsymbol{Q})$ to be a so-called "**Bregman divergence**". The following structure falls into place:

- Define $S(\boldsymbol{P}, \boldsymbol{Q}) = \boldsymbol{E}_{\boldsymbol{P}}[S(Y, \boldsymbol{Q})]$. Then $S(\boldsymbol{P}, \boldsymbol{P}) = H(\boldsymbol{P})$ due to (A).
- The term $S(Y, \boldsymbol{Q})$ is a so-called "**strict proper scoring rule**", characterized by $S(\boldsymbol{P}, \boldsymbol{Q}) \geq S(\boldsymbol{P}, \boldsymbol{P})$, with equality iff $\boldsymbol{P} = \boldsymbol{Q}$. This is a direct translation of (A) applied to $D(\boldsymbol{P}, \boldsymbol{Q}) = S(\boldsymbol{P}, \boldsymbol{Q}) - S(\boldsymbol{P}, \boldsymbol{P})$.
- The term $H(\boldsymbol{P})$ is an "**entropy**" as it is a strictly concave functional of $\boldsymbol{P}$. Its upper tangent at tangent point $\boldsymbol{Q}$ is $\boldsymbol{P} \mapsto S(\boldsymbol{P}, \boldsymbol{Q})$ due to (A). Also, (A) excludes tangent points other than $\boldsymbol{Q}$, hence renders $H(\boldsymbol{P})$ strictly concave.

Strict proper scoring rules $S(y, \boldsymbol{Q})$ generalize negative log-likelihoods.

*A.2.2 Examples of Proper Scoring Rules — Density Power Divergences:* A one-parameter family of strict proper scoring rules is as follows:

$$S_\alpha(y, \boldsymbol{Q}) = \begin{cases} -q^\alpha(y)/\alpha + \int q^{1+\alpha}\, d\mu \,/(1+\alpha) & \text{for} \quad \alpha \neq 0, -1, \\ -\log(q(y)) & \text{for} \quad \alpha = 0, \\ 1/q(y) + \int \log(q)\, d\mu & \text{for} \quad \alpha = -1. \end{cases}$$

These include proper scoring rules derived from the "density power divergences" of Basu et al. (1998) for $\alpha > 0$, the negative log-likelihood for $\alpha = 0$, and a proper scoring rule derived from the Itakura-Saito divergence for $\alpha = -1$. The two logarithmic cases ($\alpha = 0, 1$) form smooth fill-in in the manner of the logarithm in the Box-Cox family of power transforms, which makes the family well-defined for

all $\alpha \in \mathbb{R}$. The case $\alpha = 1$ corresponds to the $L_2$ distance $D_2(\boldsymbol{P}, \boldsymbol{Q}) = \int (p-q)^2 d\mu$; its proper scoring rule is $S(y, \boldsymbol{Q}) = -q(y) + \int q^2 \, d\mu/2$ and its entropy is the Gini index $H(\boldsymbol{P}) = -\int p^2 \, d\mu/2$. The power $\alpha$ is a robustness parameter, in the meaning of insensitivity to tails: robustness is gained for $\alpha \uparrow$ and sensitivity to tail probabilities for $\alpha \downarrow$. Basu et al. (1998) show that for $\alpha > 0$ the influence function is redescending for the minimum divergence estimator of the normal working model. For $\alpha \leq -1$ the divergence is so sensitive to small probabilities (hence the opposite of robust) that model densities $q(y)$ need to have tails lighter even than normal distributions.

*A.2.3 Proper Scoring Rules for Regression:* When applying a proper scoring rule $S(y, \boldsymbol{Q})$ to regression, scoring is on the conditional response distributions $\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}};\boldsymbol{\theta}} = \boldsymbol{Q}(dy|\vec{\boldsymbol{x}};\boldsymbol{\theta})$ in light of a response value $y$ at $\vec{\boldsymbol{x}}$. The resulting objective function is therefore:

$$\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) \;=\; S(y, \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}};\boldsymbol{\theta}}) \,,$$

which is used to construct a regression functional with argument $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ by

$$\boldsymbol{\theta}(\boldsymbol{P}) = \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \boldsymbol{E}_{\boldsymbol{P}}[\mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] \,.$$

Fisher consistency follows from the fact that if $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}} = \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}};\boldsymbol{\theta}_0}$, then $\boldsymbol{\theta}_0$ minimizes the objective function conditionally at each $\vec{\boldsymbol{x}}$ due to proper scoring:

$$\boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}}[\mathcal{L}(\boldsymbol{\theta}_0; Y, \vec{\boldsymbol{x}})] \leq \boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}}[\mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{x}})] \quad \forall \boldsymbol{\theta}, \; \forall \vec{\boldsymbol{x}}. \tag{14}$$

The same holds after averaging over arbitrary regressor distributions $\boldsymbol{P}_{\vec{\boldsymbol{X}}}(d\vec{\boldsymbol{x}})$:

$$\boldsymbol{E}_{\boldsymbol{P}}[\mathcal{L}(\boldsymbol{\theta}_0; Y, \vec{\boldsymbol{X}})] \leq \boldsymbol{E}_{\boldsymbol{P}}[\mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] \quad \forall \boldsymbol{\theta},$$

and hence $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{\theta}_0$.

*A.2.4 Pointwise Bregman Divergences from Convex Functions:* We illustrate one simple way of constructing what one may call "pointwise" Bregman divergences to convey the role of convex geometry. (We use here convex rather than concave functions, but this is immaterial for the construction.) If $\phi(q)$ is a strictly convex smooth function, define the associated discrepancy between two values $p$ and $q$ (in this order) to be $d(p, q) = \phi(p) - (\phi(q) + \phi'(q)(p - q))$. The term in parens is the subtangent of $\phi(\cdot)$ at $q$ as a function of $p$, hence $d(p, q) \geq 0$ holds due to convexity, and $d(p, q) = 0$ iff $p = q$ due to strict convexity. Note $d(p, q)$ is *not* generally symmetric in its arguments. The associated Bregman divergence between distributions $\boldsymbol{P}$ and $\boldsymbol{Q}$ is obtained by applying $d(p, q)$ to the respective densities $p(y)$ and $q(y)$, integrated wrt the dominating measure $\nu(dy)$:

$$\begin{aligned} D(\boldsymbol{P}, \boldsymbol{Q}) \;&=\; \int \phi(p(y))\nu(dy) - \int \phi(q(y))\nu(dy) - \int \phi'(q(y))(p(y) - q(y))\nu(dy) \\ &=\; -H(\boldsymbol{P}) + H(\boldsymbol{Q}) - \boldsymbol{E}_{\boldsymbol{P}}[\phi'(q(Y))] + \boldsymbol{E}_{\boldsymbol{Q}}[\phi'(q(Y))], \end{aligned}$$

where $H(\boldsymbol{Q}) = -\int \phi(q(y))\nu(dy)$ is the associated entropy and

$$S(y, \boldsymbol{Q}) = -\phi'(q(y)) + \boldsymbol{E}_{\boldsymbol{Q}}[\phi'(q(Y))] + H(\boldsymbol{Q}).$$

Special cases: K-L divergence for $\phi(q) = q \log(q)$; $L_2$ distance for $\phi(q) = q^2$.

*A.2.5 Density Power Divergences in Greater Detail:* Applying the preceding subsection to power transformations, suitably transformed to convexity following the Box-Cox transformation scheme, one obtains the family of density power divergences. The following is a one-parameter family of convex functions defined *for all* $\alpha \in \mathbb{R}$:

$$\phi_\alpha(q) = \begin{cases} q^{1+\alpha}/(\alpha(1+\alpha)) - q/\alpha \ + \ 1/(1+\alpha) & \text{for} \quad \alpha \neq 0, -1, \\ q\log(q) - q + 1 & \text{for} \quad \alpha = 0, \\ -\log(q) + q - 1 & \text{for} \quad \alpha = -1, \end{cases}$$

The linear terms in $q$ and the constants are irrelevant but useful to normalize $\phi_\alpha(1) = 0$ and $\phi'_\alpha(1) = 0$ for all $\alpha \in \mathbb{R}$ and to achieve the logarithmic limits for $\alpha = 0$ and $\alpha = -1$. The derivatives are:

$$\phi'_\alpha(q) = \begin{cases} q^\alpha/\alpha - 1/\alpha & \text{for} \quad \alpha \neq 0, -1, \\ \log(q) & \text{for} \quad \alpha = 0, \\ -1/q + 1 & \text{for} \quad \alpha = -1, \end{cases}$$

The associated Bregman discrepancies are:

$$d_\alpha(p, q) = \begin{cases} p^{1+\alpha}/(\alpha(1+\alpha)) \ + \ q^{1+\alpha}/(1+\alpha) - pq^\alpha/\alpha & \text{for} \quad \alpha \neq 0, -1, \\ p\log(p/q) + q - p & \text{for} \quad \alpha = 0, \\ -\log(p/q) + p/q & \text{for} \quad \alpha = -1, \end{cases}$$

Integrated to form Bregman divergences for pairs of densities $p = p(y)$ and $q = q(y)$ of $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively, one obtains:

$$D_\alpha(\boldsymbol{P}, \boldsymbol{Q}) = \begin{cases} \int \left(p^{1+\alpha}/(\alpha(1+\alpha)) \ + \ q^{1+\alpha}/(1+\alpha) - pq^\alpha/\alpha\right) d\mu) & \text{for} \quad \alpha \neq 0, -1, \\ \int p\log(p/q)\, d\mu & \text{for} \quad \alpha = 0, \\ -\int (\log(p/q) + p/q)\, d\mu & \text{for} \quad \alpha = -1, \end{cases}$$

The proper scoring rules associated with density power divergences (neglecting constants) are as follows:

$$S_\alpha(y, \boldsymbol{Q}) = \begin{cases} -q^\alpha(y)/\alpha + \int q^{1+\alpha}\, d\mu \,/(1+\alpha) & \text{for} \quad \alpha \neq 0, -1, \\ -\log(q(y)) & \text{for} \quad \alpha = 0, \\ 1/q(y) + \int \log(q)\, d\mu & \text{for} \quad \alpha = -1. \end{cases}$$

The associated entropies are as follows:

$$H_\alpha(\boldsymbol{Q}) = \begin{cases} -\int q^{1+\alpha}\, d\mu/(\alpha(1+\alpha)) & \text{for} \quad \alpha \neq 0, -1, \\ -\int q\log(q)\, d\mu & \text{for} \quad \alpha = 0, \\ \int \log(q)\, d\mu & \text{for} \quad \alpha = -1. \end{cases}$$

## A.3 Partial Influence Functions with regard to Regressor Distributions

Remark on notation: We have a need to explicitly note the distribution at which the influence function is created. Recall the definition from Section 6.6:

$$\boldsymbol{IF}(y, \vec{\boldsymbol{x}}; \boldsymbol{P}) \ = \ \frac{d}{dt}\Big|_{t=0} \boldsymbol{\theta}((1-t)\boldsymbol{P} + t\delta_{(y,\vec{\boldsymbol{x}})}).$$

This definition can be mapped to the interpretation of regression functionals as having two separate arguments, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ by splitting the pointmass $\delta_{(y,\vec{\boldsymbol{x}})}$

off to the two arguments: The conditional response distribution is $(1-t)\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}} + t\delta_y$ at this particular $\vec{\boldsymbol{x}}$, leaving those at all other $\vec{\boldsymbol{x}}'$ unchanged; the regressor distribution is changed to $(1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\delta_{\vec{\boldsymbol{x}}}$.

We show that the partial influence functions wrt $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is a shown in Proposition 6.6. We start with the integrated form of the derivative:

$$\boldsymbol{IF}(\boldsymbol{P}';\boldsymbol{P}) \;=\; \frac{d}{dt}\big|_{t=0}\,\boldsymbol{\theta}((1-t)\boldsymbol{P}+t\boldsymbol{P}') \;=\; \int \boldsymbol{IF}(y,\vec{\boldsymbol{x}};\boldsymbol{P})\,\boldsymbol{P}'(dy,d\vec{\boldsymbol{x}}).$$

which uses the fact that $\int \boldsymbol{IF}(Y,\vec{\boldsymbol{X}};\boldsymbol{P})d\boldsymbol{P} = \boldsymbol{0}$. To form the partial influence function wrt $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ holding $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ fixed, we rewrite the expansion with $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ being the same for $\boldsymbol{P}'$ and $\boldsymbol{P}$:

$$\frac{d}{dt}\big|_{t=0}\,\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes ((1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}' + t\boldsymbol{P}_{\vec{\boldsymbol{X}}})) \;=\; \int\int \boldsymbol{IF}(y,\vec{\boldsymbol{x}})\,\boldsymbol{P}(dy|d\vec{\boldsymbol{x}})\boldsymbol{P}'(d\vec{\boldsymbol{x}}), \tag{15}$$

which shows that the partial influence function wrt $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is

$$\boldsymbol{IF}(\vec{\boldsymbol{x}};\boldsymbol{P}_{\vec{\boldsymbol{X}}}) \;=\; \boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}};\boldsymbol{P})|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}].$$

(We assumed that if $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is an acceptable regressor distribution, so is a mixture $(1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\delta_{\vec{\boldsymbol{x}}}$ for small $t>0$ and any $\vec{\boldsymbol{x}}$.)

To show Proposition 6.6, if we have well-specification, then $\boldsymbol{\theta}((1-t)\boldsymbol{P}+t\delta_{\vec{\boldsymbol{x}}}) = \boldsymbol{\theta}(\boldsymbol{P})$, hence $\boldsymbol{IF}(\vec{\boldsymbol{x}};\boldsymbol{P}) = 0$. For the converse, we use the following integral representation, which is integrating up derivatives along a convex segment:

$$\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}{}') \;=\; \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) + \int \boldsymbol{IF}(\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}';(1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}')\,dt.$$

As a consequence, if $\boldsymbol{IF}(\vec{\boldsymbol{x}};(1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}') = 0$ for all $\vec{\boldsymbol{x}}$ at all regressor distributions, then $\boldsymbol{IF}(\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}';(1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}') = 0$ for all $\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}'$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, hence $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}{}') = \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}})$ for all $\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}'$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. □

## A.4 Proof of Lemma 3.3.3

The "if" part is trivial as it involves taking expectations wrt arbitrary $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. The "only if" part follows by observing that for any acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ with $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}_0$ there must exist $\vec{\boldsymbol{x}}$ for which $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}_0;Y,\vec{\boldsymbol{X}})|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}] \neq \boldsymbol{0}$. Mixtures $\boldsymbol{P}^w_{\vec{\boldsymbol{X}}} = (1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\delta_{\vec{\boldsymbol{x}}}$ for $0<t<1$ will then also be acceptable (see Section A.1), but they will not satisfy the EE for $\boldsymbol{\theta}_0$, hence $\boldsymbol{\theta}(\cdot)$ is not independent of the regressor distribution for this conditional response distribution. □

## A.5 Reweighting and Partially Additive Models

We discuss the type of diagnostic illustrated in Section 5.2 where slopes $\beta_j(\boldsymbol{P})$ are reweighted wrt to their own regressors, $Z = X_j$.

If we interpret the slope $\beta_j(\boldsymbol{P})$ of the regressor $X_j$ as a partial derivative of the best approximation to the response surface, then by localizing with a weight function $w_\xi(X_j)$ we heuristically interpret $\beta_j(w_\xi(X_j)\boldsymbol{P})$ as partial derivative of the best approximation conditional on $X_j \approx \xi$. If this localized partial derivative

is not constant in $\xi$, it indicates some form of nonlinearity of the response surface as a function of $X_j$ (linearly adjusted for, and averaged over, all other regressors).

This line of thinking may suggest that this diagnostic may be related to a partially additive regression of the form

$$Y \approx s(X_j) + \sum_{k(\neq j)} \beta_k X_k, \tag{16}$$

where $s(\cdot)$ is a smooth function of $X_j$. The heuristic correspondence with the diagnostic is this:

$$s'(\xi) \approx \beta_j(w_\xi(X_j)\boldsymbol{P}).$$

Note that in the diagnostic all adjustments are *linear* in the other regressors, not nonlinearly additive, hence the diagnostic does not correspond to a full additive model $Y \approx \sum_k s_k(X_k)$, which adjusts $s_j(X_j)$ for all other $s_k(X_k)$.

Furthermore, there exists even a difference between the diagnostic and the partially additive regression (16):

- The diagnostic adjusts $\beta_j(\cdot)$ linearly for all other $X_k$ *with regard to the reweighted distribution* $w_\xi(X_j)\boldsymbol{P}$, whereas
- the partially additive regression (16) adjusts $s(X_j)$ linearly for all other $X_k$ *with regard to the raw distribution* $\boldsymbol{P}$.

These subtle differences are to be considered when surprising effects are observed under reweighting.

### A.6 A Connection of Reweighting to Nonparametrics

Generalizing the idea of reweighting and using many — possibly infinitely many — weight functions provides a natural bridge from parametric to nonparametric regression, namely, by using reweighting functions that are "running Parzen kernels" as, for example, in local linear smoothing. The following are a few steps to describe the general idea of localizing regression functionals when the regressor space is $\mathbb{R}^p$: Let $\tilde{w}_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}})$ be a family of Parzen kernels, each member centered at a location $\vec{\boldsymbol{\xi}}$ in regressor space, an example being Gaussian kernels $\tilde{w}_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}}) \sim \exp(-\|\vec{\boldsymbol{x}} - \vec{\boldsymbol{\xi}}\|^2/(2\sigma^2))$. Then $w_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}}) = \tilde{w}_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}})/\boldsymbol{E}_{\boldsymbol{P}}[w_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{X}})]$ is a weight function that is normalized for $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ at each $\vec{\boldsymbol{\xi}}$. Finally obtain the value of the regression functional localized at $\vec{\boldsymbol{\xi}}$:

$$\boldsymbol{\theta}_{\vec{\boldsymbol{\xi}}}(\boldsymbol{P}) = \boldsymbol{\theta}(\, w_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{X}})\, \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}\,). \tag{17}$$

Two special cases:

- If $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{E}_{\boldsymbol{P}}[Y]$, then $\vec{\boldsymbol{\xi}} \mapsto \boldsymbol{\theta}_{\vec{\boldsymbol{\xi}}}(\boldsymbol{P})$ is a regularized approximation to the response surface $\vec{\boldsymbol{\xi}} \mapsto \boldsymbol{E}_{\boldsymbol{P}}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{\xi}}]$, the result of local averaging.
- If $\boldsymbol{\theta}(\cdot)$ is the linear OLS functional, then $\boldsymbol{\theta}_{\vec{\boldsymbol{\xi}}}(\boldsymbol{P})$ consists of a local intercept and local slopes at each location $\vec{\boldsymbol{\xi}}$, the latter forming a regularized approximation to the gradient at $\vec{\boldsymbol{\xi}}$. If we define $f(\vec{\boldsymbol{\xi}}) = \boldsymbol{\theta}_{\vec{\boldsymbol{\xi}}}(\boldsymbol{P})'\vec{\boldsymbol{\xi}}$, then $f(\vec{\boldsymbol{\xi}})$ is a locally linear approximation to the response surface $\vec{\boldsymbol{\xi}} \mapsto \boldsymbol{E}_{\boldsymbol{P}}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{\xi}}]$.

Estimating smooth functions and comparing them to linear ones has been a diagnostic idea for some time, and a particularly useful approach along these lines is by fitting additive models (Hastie and Tibshirani, 1990). In the next subsection we will pursue a different diagnostic idea that stays closer to the regression functional of interest.