Models as Approximations — Part II: A General Theory of Model-Robust Regression

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

Wharton – University of Pennsylvania‡

Abstract. We discuss a model-robust theory for general types of regression in the simplest case of iid observations. The theory replaces the parameters of parametric models with statistical functionals, to be called "regression functionals" and defined on large non-parametric classes of joint x-y distributions without assuming a working model. Examples of regression functionals are the slopes of OLS linear equations at largely arbitrary x-y distributions (see Part I). More generally, regression functionals can be defined by minimizing objective functions or solving estimating equations at joint x-y distributions. The role of parametric models is reduced to heuristics for generating objective functions and estimating equations without assuming them as correct. In this framework it is possible to achieve the following: (1) explicate the meaning of mis/well-specification for regression functionals, (2) decompose sampling variability into two components, one due to the conditional response distributions and another due to the regressor distribution interacting (conspiring) with misspecification, (3) exhibit plug-in (and hence sandwich) estimators of standard error as limiting cases of x-y bootstrap estimators.

AMS 2000 subject classifications: Primary 62J05, 62J20, 62F40; secondary 62F35, 62A10.
Key words and phrases: Ancillarity of regressors, Misspecification, Econometrics, Sandwich estimator, Bootstrap.

"The hallmark of good science is that it uses models and 'theory' but never believes them." (J.W. Tukey, cited by D. Brillinger, Toronto, 2016)

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.
[†]Supported in part by NSF Grant DMS-10-07689.

1. INTRODUCTION

We extend the model-robust framework of Part I from linear OLS to arbitrary parametric regression methods based on iid x-y observations. The point of view is that regression models are approximations and should not be thought of as generative truths. Because parameters of a working model define quantities of interest only under conditions of model correctness, it is necessary to extend the parameters beyond the model. In doing so one arrives at the functional point of view of regression, where parameters are not merely labels for the distributions in the model; instead, they are statistical functionals defined on quite arbitrary joint x-y distributions. They will be called "***regression functionals***."

The move from traditional regression parameters to regression functionals does not imply that every functional produces meaningful values on every distribution. It is rather held that the problem of defining a quantity of potential interest should be separated from the question of its meaningfulness. Having a quantity defined as a regression functional on a broad class of distributions creates the freedom to decide on which distributions the quantity is insightful. In actual data analysis it is possible that diagnostics reveal a model to be misspecified in some ways, and yet one may want to retain it for its interpretability.

Regression models are chosen to a large extent based on the measurement type of the response (including: interval scale, ratio scale, counting, binary, multi-categorical, spatial, multi-response), as when choosing a suitable generalized linear model. The model-robust regression framework described here applies to all response types as long as the observations are iid (we hence exclude correlated data such as time series). If the specifics of the regression situation are inessential for the framework, its essentials are as follows:

- The variables are divided into regressor and response variables, denoted $\vec{\boldsymbol{X}}$ and Y, respectively (the typographic distinction being a hold-over from Part I where regressors are multi-dimensional and responses uni-dimensional, an assumption that is no longer made).
- The association between $\vec{\boldsymbol{X}}$ and Y is described asymmetrically by focusing on the conditional response distribution $Y|\vec{\boldsymbol{X}}$ while attempting to ignore the marginal distribution of the regressors $\vec{\boldsymbol{X}}$, thought to be "ancillary".

In parametric regression modeling one assumes a parametrized family of conditional response distributions $\{\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ or, in terms of densities, $\{q(y|\vec{\boldsymbol{x}};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. The model-robust point of view is not to trust such models but use them pragmatically to construct regression functionals.

Because regression functionals are at the core of the framework, a first goal here is to catalogue some of the ways of constructing them: (1) optimization of objective functions, in particular maximum likelihood based on a working model, and (2) solving estimating equations, in particular the score equations of a working model. Thus the two ways have strong overlap, but it is the literature on estimating equations where the functional point of view of parameters is more firmly embedded. This first part of the article is purely tutorial and not original, but a necessity in light of what follows.

The second and most original goal of this article is to introduce a notion of mis- and well-specification for regression functionals. While it is clear what this notion means in the context of parametric models, it is less clear that a related notion of mis/well-specification can be defined for regression functionals. Here are some thoughts based on linear OLS that should make the idea plausible:

- Consider the slopes in a multiple linear regression and ask when they are most meaningful. The "obvious" answer would refer to well-specification of the model, $Y \sim \mathcal{N}(\vec{\boldsymbol{X}}'\boldsymbol{\beta}, \sigma^2)$, but this is too strong: homoskedasticity and Gaussianity are not immediately related to slopes which describe linearity of the conditional response mean. Hence a weaker condition is more plausible: the slopes $\boldsymbol{\beta}$ are perfectly meaningful if the the conditional mean function is linear, $\boldsymbol{E}[Y|\vec{\boldsymbol{X}}] = \boldsymbol{\beta}'\vec{\boldsymbol{X}}$. Equivalently, the gradient $\nabla_{\vec{\boldsymbol{x}}}\boldsymbol{E}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \boldsymbol{\beta}$ exists and is constant.
- Next consider the average noise variance, measured by a regression functional defined as $\sigma^2(\boldsymbol{P}) = \boldsymbol{E}[\boldsymbol{V}[Y|\vec{\boldsymbol{X}}]]$, and ask when it is most meaningful. Again the obvious answer is "under model well-specification", but a weaker and more plausible condition is homoskedasticity: $\boldsymbol{V}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \sigma^2$ for (almost) all $\vec{\boldsymbol{x}}$.

These examples have a shared feature: In the scenarios where these quantities are most meaningful, they are constant across regressor space. A constant gradient implies linearity and a constant conditional response variance means homoskedasticity. Such constancy, if it holds, confers special meaning to the value of the functional because it characterizes the conditional response distribution across all locations in regressor space; it is not some compromise of different values in different locations. It is therefore sensible to make this condition of *constancy across regressor space* the basis of what it means to say a regression functional is well-specified for a particular data distribution.

The condition requires an important modification before it reaches its final form, and the remaining issue can again be illustrated with slope functionals: a slope at a point $\vec{\boldsymbol{x}}$ is mathematically a partial derivative and requires knowledge of $\boldsymbol{E}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}]$ in a neighborhood of $\vec{\boldsymbol{x}}$. Statistically, the issue can be resolved by replacing locations $\vec{\boldsymbol{x}}$ with (non-collinear) regressor distributions. The question of constancy across regressor space can therefore be recast as *constancy across (acceptable) regressor distributions.* With this final step we have arrived at a coherent definition of well-specification for regression functionals which we state informally as follows:

> **Definition:** A regression functional is well-specified for a conditional response distribution of $Y|\vec{\boldsymbol{X}}$ if the functional does *not* depend on the marginal regressor distribution of $\vec{\boldsymbol{X}}$.

Intuitively the meaning of this notion of well-specification is that the target of estimation is independent of where in regressor space the response is observed. This being an ideal, in the world of real data where misspecification is the rule rather than the exception, it should be expected that the target of estimation *does* depend on where in regressor space the response is observed.

Leveraging this notion of mis/well-specification for regression functionals, a third goal of this article is to extend to all types of regression the important fact (described in Part I) that, under misspecification, sampling variability of linear OLS estimates has *two* sources:

- the noise $\epsilon = Y - \boldsymbol{E}[Y|\vec{\boldsymbol{X}}]$;

- the nonlinearity $\eta(\vec{\boldsymbol{X}}) = \boldsymbol{E}[Y|\vec{\boldsymbol{X}}] - \boldsymbol{\beta}'\vec{\boldsymbol{X}}$.

A difficulty in attempting a generalization to arbitrary regression functionals and their plug-in estimates is that the notions of noise and nonlinearity are not available in all types of regressions. However, the following formulation proves to be universally applicable: The sampling variability of plug-in estimates of all regression functionals has two sources,

- the first source being $Y|\vec{\boldsymbol{X}}$, the $\mathbf{X}$-conditional randomness of the response Y;
- the second source being the randomness of $\mathbf{X}$ in the presence of misspecification.

Here a fortuitous feature of our definition of mis/well-specification comes to light: If the regression functional is well-specified for $Y|\vec{\boldsymbol{X}}$, its value is the same for all $\mathbf{X}$ distributions, including the empirical $\mathbf{X}$ distributions of observed data. As a result, under well-specification there does not exist sampling variability due to $\vec{\boldsymbol{X}}$. Conversely, if there is misspecification, the functional will vary between empirical $\mathbf{X}$ distributions and hence generate sampling variability in estimates. What was a "conspiracy of nonlinearity and random-$\mathbf{X}$" for linear OLS in Part I is now a "conspiracy of misspecification and random $\mathbf{X}$", where "misspecification" has the above technical meaning for regression functionals. — As in Part I for linear OLS, we will describe three CLTs, one for the full sampling variability and one each for the components due to $Y|\vec{\boldsymbol{X}}$ and $\vec{\boldsymbol{X}}$, respectively.

A fourth and final goal is to generalize a result of Part I that links two types of model-robust frequentist inference: the *x-y* bootstrap (as opposed to the residual bootstrap) and the plug-in method applied to asymptotic variance formulas (which usually results in sandwich estimators). To this end we require again a generalized version of the bootstrap whereby the resample size M can differ from the sample size N. It will be seen that the argument of Part I generalizes in full: plug-in estimators are always limits of the M-of-N bootstrap as $M \to \infty$.

The article proceeds as follows: Section 2 describes some of the ways in which regression functionals can be defined. Section 3 introduces the notion of mis/well-specification for a regression functional on a given conditional response distribution. Section 4 explains the canonical decomposition of sampling variability under misspecification for plug-in estimates of regression functionals. Section 5 recounts relevant model-robust CLTs for plug-in estimates of regression functionals. Section 6, finally, describes the connection between plug-in estimators and bootstrap estimators of standard error.

Remark 1: Model robustness is not the same as classical outlier/heavy-tail robustness. In regression the latter generally questions the assumed error distribution but not the fitted equation. As presented here, model robustness questions both and retains only the (equally questionable) assumption of iid sampling. Some types of model-robust theory that relax the iid assumption by allowing time series structure are common in econometrics (e.g., White 1994).

Remark 2: The idea of model robustness has a long history that includes Huber (1967), Kent (1982), and a long tradition of theorizing about misspecification starting with White (1980a, 1980b, 1981, 1982). While the idea that models are not "truths" is widely accepted, it is necessary to understand that the acceptance of "misspecification" and the view of "models as approximations" has real consequences for the targets of estimation, the sources of randomness in estimation, and the types of inference that are valid under misspecification.

2. TARGETS OF ESTIMATION: REGRESSION FUNCTIONALS

This section first lays out some assumptions about the universe of distributions on which a regression functional is defined. The section subsequently describes two ways of constructing regression functionals.

2.1 Preliminaries on Joint Distributions

The population description of regression based on observational data is in terms of two random variables that take on values in respective measurable spaces: the regressor $\vec{\boldsymbol{X}}$: $\boldsymbol{\Omega} \to \mathcal{X}$ and the response Y: $\boldsymbol{\Omega} \to \mathcal{Y}$ with a joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$, a conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and a regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. We may express the connection between them using the notation

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

As mentioned earlier, the typographic distinction between $\vec{\boldsymbol{X}}$ and Y is a hold-over from the OLS context of Part I where $\vec{\boldsymbol{X}}$ denoted a random vector of quantitative regressor variables and Y a quantitative response variable. This is no longer assumed and the regressor and response spaces $\mathcal{X}$ and $\mathcal{Y}$ are now entirely arbitrary.

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:

- Expectations will be assumed to exist for all distributions in $\mathcal{P}$.
- If the 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 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.
- With the preceding bullet 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 $\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}'$ will also be acceptable. The reason is that mixing can only enlarge but not diminish the support of the distribution, hence identifiability of fitted parameters will be inherited from $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ irrespective of $\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}'$.
- The preceding bullet ensures that the set of acceptable regressor distributions is rich, in fact so rich that $E_{\boldsymbol{P}_{\vec{\boldsymbol{X}}}}[f(\vec{\boldsymbol{X}})] = 0$ for all acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ entails $f \equiv 0$.
- For working models $\{\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}: \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ (not treated as well-specified) 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 $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.
- Where conditional model densities $q(y|\vec{\boldsymbol{x}};\boldsymbol{\theta})$ of the response appear, they will be with regard to some dominating measure $\nu(dy)$.
- For plug-in estimation it will be required that empirical distributions $\hat{\boldsymbol{P}} = \hat{\boldsymbol{P}}_N$ arising as N iid draws from $\boldsymbol{P} \in \mathcal{P}$ will be in $\mathcal{P}$ with probability 1. This requirement can be limited to sufficiently large N. For example, one needs $N \geq p+1$ non-collinear observations in order to fit a linear equation with intercept using OLS on a p-dimensional regressor space.

- 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 $0 < t < 1$.
- $\mathcal{P}$ will be assumed to be convex, hence closed under finite mixtures.

2.2 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 chosen criterion. This approach can be generalized to other objective functions $\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$:

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

A common choice for $\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$ is the negative log-likelihood of a parametric regression model for $Y|\vec{\boldsymbol{X}}$ given by conditional response distributions $\{\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with conditional densities $\{q(y \,|\, \vec{\boldsymbol{x}}; \boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\theta}$ has the traditional meaning of a parameter. Not assuming the model to be well-specified, its only purpose is to serve as a heuristic to suggest a meaningful objective function:

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

The resulting regression functional will be called a ML functional. (For technical conditions under which it is well-defined see White (1982) and Huber (1967).) ML functionals are fundamental because they naturally extend traditional parameters of major classes of regression models such as GLMs. Technically, they also comprise M-estimators based on Huber ρ functions (Huber 1964), including least absolute deviation (LAD) or L_1 objective functions for conditional medians, and tilted L_1 versions for arbitrary conditional quantiles. All of these objective functions can be interpreted as negative log-likelihoods of certain distributions, even if they may not usually be interpreted as 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.

A natural extension of ML functionals is in terms of so-called "proper scoring rules", the subject of Appendix A. The regression functionals that result may be called "proper scoring or PS functionals", a superset of ML functionals. Proper scoring rules can be used to form discrepancy measures between a conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and model distributions $\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}$, generalizing the expected negative log-likelihood of the conditional response distribution with regard to the model, $\boldsymbol{E}[-\log q(Y|\vec{\boldsymbol{x}}; \boldsymbol{\theta}) \,|\, \vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}]$. The discrepancy measure between $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}$ is then averaged with regard to the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, which in the case of the negative log-likelihood results in $\boldsymbol{E}_{\boldsymbol{P}}[-\log q(Y|\vec{\boldsymbol{X}}; \boldsymbol{\theta})]$. This average discrepancy is the crition whose minimization with regard to $\boldsymbol{\theta}$ defines the value of the PS functional $\boldsymbol{\theta}(\boldsymbol{P})$ for the data distribution $\boldsymbol{P}$. PS and hence ML functionals have the important property of Fisher consistency:

$$\boldsymbol{P} = \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}_0} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}} \quad\Rightarrow\quad \boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{\theta}_0. \tag{1}$$

See Appendix A for the connections of PS functionals to Bregman divergences and outlier/heavy-tail robustness.

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

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) \;:=\; \mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) + \lambda \mathcal{R}(\boldsymbol{\theta}). \tag{2}$$

Special cases are Ridge and Lasso penalties. Note that (2) results in one-parameter families of penalized functionals $\boldsymbol{\theta}_\lambda(\boldsymbol{P})$ defined for populations as well, whereas in practice penalty parameters are thought to apply to finite data with $\lambda_N \to 0$.

2.3 Regression Functionals from Estimating Equations

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

$$\boldsymbol{E}_{\boldsymbol{P}}[\psi(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] \;=\; \mathbf{0}. \tag{3}$$

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

$$\boldsymbol{\psi}_{OLS}(\boldsymbol{\beta}; y, \vec{\boldsymbol{x}}) \;=\; \vec{\boldsymbol{x}} y - \vec{\boldsymbol{x}} \vec{\boldsymbol{x}}' \,\boldsymbol{\beta} \;=\; \vec{\boldsymbol{x}}(y - \vec{\boldsymbol{x}}' \,\boldsymbol{\beta}). \tag{4}$$

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 important extension is to situations where the number of moment conditions (the dimension of $\boldsymbol{\psi}$) is larger than the dimension of $\boldsymbol{\theta}$. This is known as "Generalized Method of Moments" (GMM, Hansen 1982) which can be used for causal inference based on numerous instrumental variables.

Another extension is to situations where clustered data have intra-cluster dependence, known as "Generalized Estimating Equations" (GEE, Liang and Zeger 1986). This is a fixed-$\mathbf{X}$ approach that assumes well-specification of the mean function but allows misspecification of the variance and intra-cluster dependence.

2.4 The Point of View of Regression Functionals and its Implications

Theories of parametric models deal with the issue that a given model parameter has many possible estimators, as in the normal model $\mathcal{N}(\mu, \sigma^2)$ where the mean is in various ways the optimal estimate of μ whereas the median is a less efficient estimate. 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 possibly inconsistent estimators of the same parameter; rather, they represent different statistical functionals with associated plug-in estimators. Similarly, slopes obtained by linear OLS and linear LAD represent 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. Whether OLS or LAD should be used is not a matter of right or wrong but of extraneous considerations that may include differences in the degrees of misspecification in the sense of Section 3 below.

In a similar vein, 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 β_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. If $M \subset M'$, then $\beta_{j \cdot M'} - \beta_{j \cdot M}$ is not a bias but a difference between two regression functionals. If it is small in magnitude, 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 is a case of Simpson's paradox for this regressor.

It should further be noted that regression functionals depend on the full joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ of the response and the regressors, not just the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. Conventional regression parameters characterize only the conditional response distribution by assuming it to be of the form $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} = \boldsymbol{Q}_{Y|\vec{\boldsymbol{X}};\boldsymbol{\theta}}$, while the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is sidelined as ancillary. That the ancillarity argument for the regressors is not valid under conditions of misspecification was documented in Part I, Section 4. In the following Section 3 this fact will be the basis of a new concept of misspecification for regression functionals.

3. MIS-/WELL-SPECIFICATION OF REGRESSION FUNCTIONALS

The introduction motivated a notion of mis/well-specification for regression functionals, and this section provides the technical notation. Before the reader continues, he/she may want recall the intuitions given in Part I, Section 4, Proposition 4.1, and in particular Figure 2. The proposition shows for the slope functional of linear OLS that it depends on the regressor distribution if and only if the response function is nonlinear. This condition is so intriguingly generalizable that it lends itself as a master definition of mis/well-specification for arbitrary regression functionals in arbitrary types of regressions. An attractive feature of this notion is that it speaks precisely to the specific property of distributions $\boldsymbol{P}$ that is measured by the regression functional $\boldsymbol{\theta}(\boldsymbol{P})$ without commitment to a model and the peculiar distributional properties implied by it.

3.1 Definition of Well-Specification for Regression Functionals

The starting point is to decompose the joint distribution $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ into two components: the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and the marginal regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. In general, 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 $\vec{\boldsymbol{x}} \in \mathcal{X}$ (that is, we assume that a "regular version" is chosen). With these conventions, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ uniquely determine $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$, but not quite vice versa. As before we write $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} = \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}$ and reinterpret $\boldsymbol{\theta}(\cdot)$ as a functional of these two arguments:

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

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 $\boldsymbol{P}'_{\vec{\boldsymbol{X}}}$.

As indicated in Section 2.1, the term “acceptable” accounts for exclusions of regressor distributions such as non-identifiability when fitting equations, in particular perfect collinearity when fitting linear equations.

Remark: 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.
Caveat: Any notion of well-specification represents an ideal, not a reality. Well-specification in any meaning of the term is never a fact; real are only degrees of misspecification. Yet, ideals are useful because they spell out the circumstances under which a lofty goal is perfectly realized. Here the goal is that measurements of the $Y|\vec{\boldsymbol{X}}$ association be independent of where in regressor space the observations are taken.

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 is $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})]$, where $\mu(\vec{\boldsymbol{x}}) = \boldsymbol{E}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}]$. 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 $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}_0$. The reverse is true, too; see Part I, Proposition 4.1.

- 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 $\boldsymbol{\beta}(\boldsymbol{P}) = (\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,] + \boldsymbol{\Omega})^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})]$. This functional is well-specified if the conditional mean is linear, $\mu(\vec{\boldsymbol{x}}) = \boldsymbol{\beta}_0{}'\,\vec{\boldsymbol{x}}$, and $\boldsymbol{\Omega} = c\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]$ for some $c \geq 0$, in which case $\boldsymbol{\beta}(\boldsymbol{P}) = 1/(1+c)\,\boldsymbol{\beta}_0$, causing uniform shrinkage.

- The point of regression functionals is that they describe the association between a response variable Y and regressor variables $\vec{\boldsymbol{X}}$. However, for a quantitative response Y, what does it mean for the functional $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{E}[\,Y]$ 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 of course a fallacy: writing $\boldsymbol{E}[Y] = \int \boldsymbol{E}[\,Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}\,]\,\boldsymbol{P}_{\vec{\boldsymbol{X}}}(d\vec{\boldsymbol{x}})$, it follows that $\boldsymbol{E}[Y]$ is independent of $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ iff the $\vec{\boldsymbol{X}}$-conditional mean is constant: $\boldsymbol{E}[\,Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}\,] = \boldsymbol{E}[Y] \;\; \forall\vec{\boldsymbol{x}}$.

- The average conditional variance functional $\sigma^2(\boldsymbol{P}) = \boldsymbol{E}[\boldsymbol{V}[Y|\vec{\boldsymbol{X}}]]$ is well-specified iff $\boldsymbol{V}[Y|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \sigma_0^2$ is constant, in which case $\sigma^2(\boldsymbol{P}) = \sigma_0^2$.

- Fitting a linear equation by minimizing least absolute deviations (i.e., the LAD or L_1 objective function) defines a regression functional that is well-specified if there exists $\boldsymbol{\beta}_0$ such that $\text{median}[\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}] = \boldsymbol{\beta}_0{}' \vec{\boldsymbol{x}}$ for all $\vec{\boldsymbol{x}}$.
- 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. The point of such functionals is to characterize the marginal regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ alone. They could not be "well-specified" unless they are trivial 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, equation (10)).

Lemma 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 $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. Furthermore, $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}_0$ for all acceptable regressor distributions $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$.*

The corollary below is a consequence of the lemma above but could have been gleaned from Fisher consistency (1), which holds irrespective of the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$:

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

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 $\boldsymbol{\theta}_0$.

Lemma 3.3.3: *The EE functional $\boldsymbol{\theta}(\cdot)$ defined by $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] = \boldsymbol{0}$ is well-specified for $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ iff there exists $\boldsymbol{\theta}_0$ such that $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}_0; Y, \vec{\boldsymbol{X}})|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \boldsymbol{0} \;\; \forall \vec{\boldsymbol{x}}$.*

The proof is in Appendix C.

3.4 Well-Specification and Influence Functions

To state the next fact, we will make heuristic use of influence functions, which will be useful for the asymptotics of regression functionals in Section 5. (For a general treatment of influence functions see Hampel et al. (1986) and Rieder (1994).) The influence function of a regression functional, when it exists, is a form of derivative on the space of probability distributions. It is intuitive to use influence functions to characterize well-specification of regression functionals because if the functional $\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}})$ 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 derivative wrt $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ vanishes. The definition of the influence function in terms of $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ is as follows:

$$(5) \qquad \boldsymbol{IF}(y, \vec{\boldsymbol{x}}) \;:=\; \left.\frac{d}{dt}\right|_{t=0} \boldsymbol{\theta}((1-t)\boldsymbol{P} + t\delta_{(y,\vec{\boldsymbol{x}})}).$$

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

$$\boldsymbol{IF}(\vec{\boldsymbol{x}}) \;:=\; \frac{d}{dt}\Big|_{t=0} \boldsymbol{\theta}\left(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes ((1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}} + t\delta_{\vec{\boldsymbol{x}}})\right) \tag{6}$$

The proof of the following lemma is in Appendix B:

Lemma 3.4: $\quad \boldsymbol{IF}(\vec{\boldsymbol{x}}) \;=\; \boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}}[\boldsymbol{IF}(Y, \vec{\boldsymbol{X}})|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}].$

We will use as notations either $\boldsymbol{IF}(\vec{\boldsymbol{x}})$ or $\boldsymbol{E}[\boldsymbol{IF}(Y, \vec{\boldsymbol{X}})|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}]$, lightening the notational burden in the latter by dropping the subscript $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ of $\boldsymbol{E}[\cdot]$.
See again Appendix B for a proof of the following:

Proposition 3.4: *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 $\;\boldsymbol{IF}(\vec{\boldsymbol{x}}) = \boldsymbol{0} \;\; \forall \vec{\boldsymbol{x}}$.*

Influence functions will prove useful for approximations in Sections 4.5 and for asymptotics in Section 5.

3.5 Reweighting and Well-Specification

Mis/well-specification of functionals relates in useful ways to reweighting of data which can be leveraged for misspecification tests as well as nonparametric extensions:

Assumptions and Notations:*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:*

$$\tilde{\boldsymbol{P}}_{Y,\vec{\boldsymbol{X}}}(d\vec{\boldsymbol{x}}) = w(\vec{\boldsymbol{x}})\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}(d\vec{\boldsymbol{x}}); \qquad \tilde{p}(y, \vec{\boldsymbol{x}}) = w(\vec{\boldsymbol{x}})p(y, \vec{\boldsymbol{x}}),$$

where $w(\vec{\boldsymbol{x}}) \geq 0$ and $\boldsymbol{E}_{\boldsymbol{P}}[w(\vec{\boldsymbol{X}})] = 1$. Shortened notation: $\tilde{\boldsymbol{P}} = w(\mathbf{X})\boldsymbol{P}$.

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

Proof: The conditional response distribution is unchanged under $\vec{\boldsymbol{x}}$-dependent reweighting: $\tilde{\boldsymbol{P}}_{Y|\vec{\boldsymbol{X}}} = \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, and a well-specified regression functional is unchanged under a change of the regressor distribution, $\tilde{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}(d\vec{\boldsymbol{x}}) = w(\vec{\boldsymbol{x}})\boldsymbol{P}_{\vec{\boldsymbol{X}}}(d\vec{\boldsymbol{x}})$. □

In fixed-$\mathbf{X}$ theories, which necessarily assume well-specification of the model, it is common currency that reweighting the data grants consistent estimation of parameters. Translated to the current framework, this fact returns as a statement of invariance of well-specified functionals under $\vec{\boldsymbol{X}}$-dependent reweighting.

Reweighting can be used to devise misspecification tests as described by White (1980a, Section 4) for linear OLS. The idea generalizes to arbitrary types of regression and regression functionals: One tests the null hypothesis that the difference $\boldsymbol{\theta}(w_1(\vec{\boldsymbol{X}})\boldsymbol{P}) - \boldsymbol{\theta}(w_2(\vec{\boldsymbol{X}})\boldsymbol{P})$ vanishes for two $\vec{\boldsymbol{x}}$-dependent weight functions, $w_1(\cdot)$

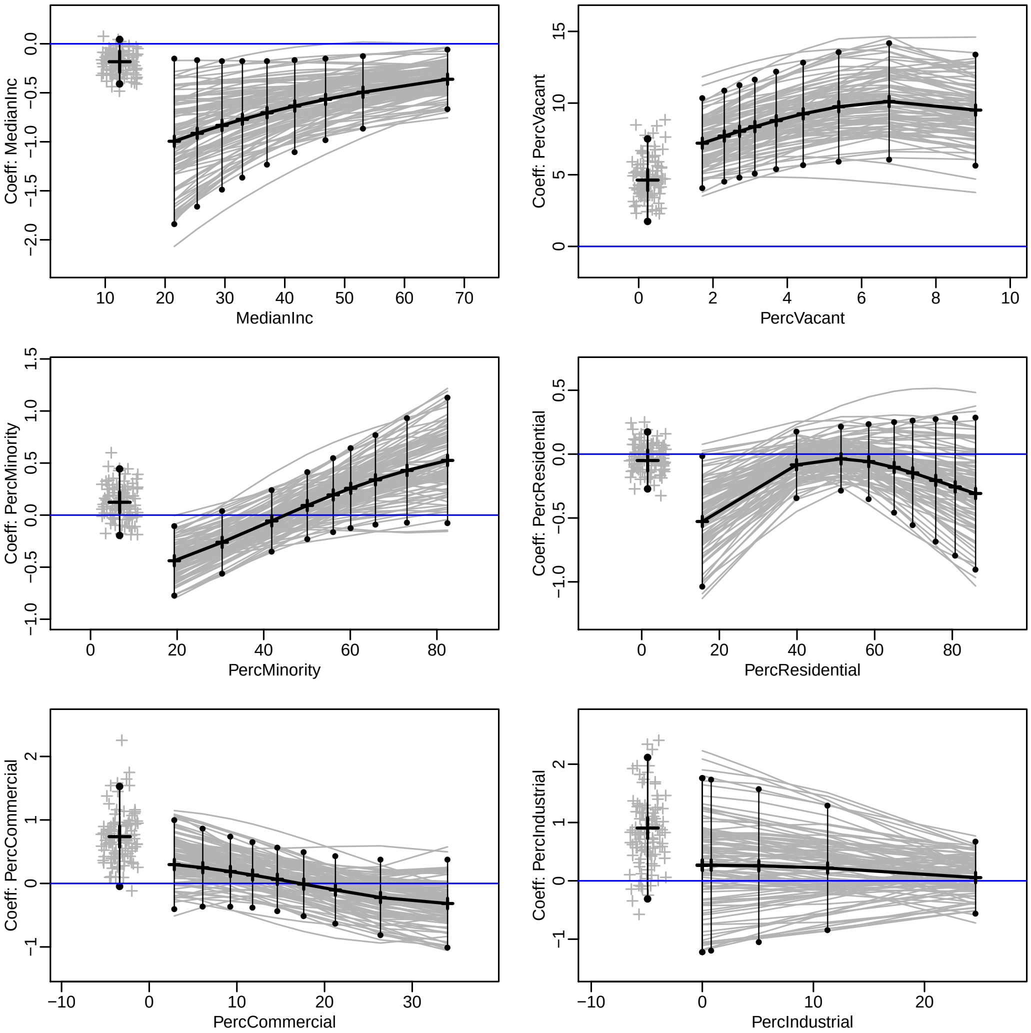

FIG 1. *Misspecification Diagnostics Based on Reweighting: How stable is a regression coefficient when the data are reweighted as a function of its regressor?*
Data: Counts of homeless in a sample of census tracts of LA as a function of various regressors; see Part I, Section 2 (Berk et al. 2008).
Vertical axis = regression coefficient
Horizontal axis = regressor
Traces = regression coefficient as a function of the reweighting center
gray traces = weighted bootstrap samples,
black trace = weighted observed data
Points on left: unweighted coefficient, not a function of the horizontal axis
gray points = unweighted bootstrap samples
black point = unweighted observed data
Vertical line segments = confidence intervals, ±2 bootstrap standard errors
The weights are Gaussian, centered (μ) at the deciles of the regressor distribution, with dispersion (σ) = 1 standard deviation of the regressor distribution.
Horizontal line = 0 level to judge statistical significance
Some Conclusions:
(1) Both `MedianInc` *and* `PercVacant` *appear more statistically significant under reweighting (see traces) than without (see points on left and compare with the 0 level line). Their traces are not flat but they stay more removed from the 0 level compared to the unweighted bootstrap values (points on left).*
(2) `PercMinority` *appears to be negatively associated with the response for low values, but positively associated for high values.*
(3) The coefficient of `PercIndustrial` *is most well-specified, but it is statistically insignificant.*

and $w_2(\cdot)$. Inference for such differences can be based on the non-parametric x-y bootstrap (see Part I, Section 8, as well as Section 6 below).

Generalizing the idea 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 $\mathrm{I\!R}^p$: Let $\tilde{w}_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}})$ be a family of Parzen kernels, each member centered in some sense at a location $\vec{\boldsymbol{\xi}}$ in regressor space, an example being a Gaussian kernels $\tilde{w}_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}}) \sim \exp(-\|\vec{\boldsymbol{x}} - \vec{\boldsymbol{\xi}}\|^2/(2\sigma^2))$. Then $w_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}}) = \tilde{w}_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{x}})/\boldsymbol{E}[w_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{X}})]$ 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 $\vec{\boldsymbol{\xi}}$:

$$\boldsymbol{\theta}_{\vec{\boldsymbol{\xi}}}(\boldsymbol{P}) := \boldsymbol{\theta}(\, w_{\vec{\boldsymbol{\xi}}}(\vec{\boldsymbol{X}})\, \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}\,). \tag{7}$$

Two special cases:

- If $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{E}[Y]$, then $\vec{\boldsymbol{\xi}} \mapsto \boldsymbol{\theta}_{\vec{\boldsymbol{\xi}}}(\boldsymbol{P})$ is a regularized approximation to the response surface $\vec{\boldsymbol{\xi}} \mapsto \boldsymbol{E}[Y|\vec{\boldsymbol{X}}\!=\!\vec{\boldsymbol{\xi}}]$, 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 $\vec{\boldsymbol{\xi}}$, which we can think of as a regularized gradient. 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 $\vec{\boldsymbol{\xi}} \mapsto \boldsymbol{E}[Y|\vec{\boldsymbol{X}}\!=\!\vec{\boldsymbol{\xi}}]$.

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.

3.6 A Model Diagnostic Based on Reweighting

The following diagnostic is directly inspired by the definition of mis/well-specification of regression functionals. The simple idea is to check the constancy of a regression functional under localized reweighting of the data. We will focus on the case where the functional is a regression coefficient associated with a particular regressor variable. In the simplest case this is the coefficient $\beta_j = \beta_j(\boldsymbol{P})$ of regressor X_j in linear OLS as layed out in Part I. Because of the analytical connection between β_j and X_j we expect that strong sensitivity of this functional to reweighting is achieved when the weights are a function of X_j. A simple example is a univariate Gaussian weight function centered at ξ:

$$\tilde{w}_\xi(x_j) \;\sim\; \exp(-(x_j - \xi)^2/(2\sigma^2))\,, \qquad w_\xi(x_j) \;=\; \tilde{w}_\xi(x_j)/\boldsymbol{E}[\tilde{w}_\xi(X_j)]\,.$$

A diagnostic is obtained by plotting β_j as a function of ξ:

$$\xi \;\mapsto\; \beta_{j,\xi}(\boldsymbol{P}) \;:=\; \beta_j(w_\xi(X_j)\,\boldsymbol{P})\,.$$

If the regression functional $\beta_j(\cdot)$ is well-specified, then $\beta_{j,\xi}(\cdot)$ is constant in ξ. Conversely, if $\beta_{j,\xi}(\cdot)$ is not constant, then $\beta_j(\cdot)$ is misspecified. This is of course only a sufficient and not a necessary criterion, but from the way the weights are constructed we expect some power from this diagnostic for detecting misspecification in the sense of Section 3.1.

To make the approach actionable, one obtains estimates $\hat{\beta}_{j,\xi} = \beta_j(\hat{w}_\xi(X_j)\hat{\boldsymbol{P}})$, where $\hat{w}_\xi(x_j)$ is a weight function that is empirically normalized to unit mass: $\hat{\boldsymbol{E}}[\hat{w}_\xi(X_j)] = 1$. For Gaussian weights centered at ξ one may use $\hat{w}_\xi(x_j) \sim \exp((x_j - \xi)^2)/(2(\alpha\hat{\sigma}_j)^2)$, where $\hat{\sigma}_j$ is the empirical standard deviation of X_j and α is a bandwidth parameter to be chosen a priori (a simple choice is $\alpha = 1$). Estimates $\hat{\beta}_{j,\xi}$ can be obtained for a grid of values $\xi \in \{\xi_1, ..., \xi_K\}$, a simple choice being the nine interior deciles of the empirical X_j distribution ($K = 9$). The resulting K values of $\hat{\beta}_{j,\xi_k}$ can be drawn as a trace over the grid values ξ_k, and they can be compared to the unweighted estimate $\hat{\beta}_j$.

The scheme just described is carried out in Figure 1 for the LA homeless data that were used in Part I, Section 2. A general issue with diagnostics is that they leave a fundamental question unanswered: Does the trace $\xi_k \mapsto \hat{\beta}_{j,\xi_k}$ vary in a statistically significant way to allow the inference that the statistical functional $\beta_j(\boldsymbol{P})$ is misspecified? To answer this question one would need direct inference for differences $\hat{\beta}_{j,\xi_k} - \hat{\beta}_{j,\xi_{k'}}$ ($k \neq k'$) as well as $\hat{\beta}_{j,\xi_k} - \hat{\beta}_j$. We found it, however, more useful to decorate the graphical traces with bootstrap replications and bootstrap confidence intervals, as shown in Figure 1. The question as to the existence of misspecification is in our experience less interesting because misspecification is nearly universal. More interesting is the question as to the nature of the misspecifications, and answering this question is greatly assisted by the bootstrap-based decors added to Figure 1. There is no need to give here a detailed discussion of the displays as the graphical details are explained in the figure caption; examples of conclusions that can be drawn from the displays are given toward the end of the caption.

A surprising general experience is that reweighted estimates $\hat{\beta}_{j,\xi_k}$ can be less *or* more statistically significant than the unweighted estimate $\hat{\beta}_j$. It can also occur that reweighted estimates $\hat{\beta}_{j,\xi_k}$ may change signs from one end of the grid (ξ_1) to the other end (ξ_K), even in a statistically significant fashion (see the regressor `PercMinority` in Figure 1). This possibility may be less surprising if the values $\hat{\beta}_{j,\xi_k}$ are interpreted as local estimates of the derivative of the X_j-marginalized response surface.

Stepping back from the minutiae of implementing a reweighting diagnostic and devising inference for it, it should be noted that the resulting mis/well-specification tests do not rely on buy-in into traditional models. The notion of well-specification and inferential diagnostics provided here concern a quantity of interest, that is, a regression functional. In linear OLS, for example, if the quantities of interest are the slopes or, for that matter, a subset of the slopes, there is no need to diagnose linearity with regard to slopes of no interest or constancy of the noise variance or normality of the noise distribution. The idea of reweighting the data is a universal device to diagnose and test well-specification of exactly the quantities of interest in most types of regression.

4. ESTIMATION OF REGRESSION FUNCTIONALS: CANONICAL DECOMPOSITION OF ESTIMATION OFFSETS

4.1 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 $(Y_i, \vec{\boldsymbol{X}}_i)$, the responses are collected in a structure $\mathbf{Y}$, and the regressors $\vec{\boldsymbol{X}}_i$ in another structure $\mathbf{X}$. We purposely avoid the terms "vector" and "matrix" because in a general theory of regression all variables — responses and regressors — can be of any type and any number. As mentioned earlier, the typographic distinctions between response and regressors are a carry-over from Part I that have no bearing anymore: the response Y_i could be vectorial, as it would be in multi-response situations, or when categorical multi-class outcomes are coded with dummy vectors for the classes. Thus both $\mathbf{Y}$ and $\mathbf{X}$ are best thought of as "data frames" in R (2008) terminology. Regression of $\mathbf{Y}$ on $\mathbf{X}$ is any attempt at describing 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 $\hat{\boldsymbol{\theta}} := \boldsymbol{\theta}(\hat{\boldsymbol{P}})$ where $\hat{\boldsymbol{P}} := (1/N)\sum \delta_{(Y_i,\vec{\boldsymbol{X}}_i)}$ is the empirical distribution. If necessary we may write $\hat{\boldsymbol{P}}_N$ for $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{\theta}}_N$ for $\hat{\boldsymbol{\theta}}$.

4.2 The Conditional Parameter of Model-Trusting Regression

This subsection generalizes the "conditional parameter" $\boldsymbol{\beta}(\mathbf{X})$ of Part I, Section 5. It will be denoted $\boldsymbol{\theta}(\mathbf{X})$. For a regression functional written as $\boldsymbol{\theta}(\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}})$ according to Section 3, the conditional parameter is obtained by partial plug-in of the empirical distribution $\hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ of the regressors, as shown in the center line of the following overview of notations:

$$\boxed{\begin{aligned} \boldsymbol{\theta}(\boldsymbol{P}) &= \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}}), \\ \boldsymbol{\theta}(\mathbf{X}) &:= \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}), \qquad \hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}} = (1/N)\textstyle\sum \delta_{\vec{\boldsymbol{X}}_i}, \\ \hat{\boldsymbol{\theta}} &:= \boldsymbol{\theta}(\hat{\boldsymbol{P}}). \end{aligned}}$$

It is legitimate to write $\boldsymbol{\theta}(\mathbf{X})$ because $\mathbf{X}$ and $\hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ contain the same information; the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ is implied and not shown. The main point is:

- In model-trusting theories that condition on $\mathbf{X}$, the target of estimation is $\boldsymbol{\theta}(\mathbf{X})$. They assume $\boldsymbol{\theta}(\mathbf{X})$ is the same for all acceptable $\mathbf{X}$.
- In model-robust theories that do not condition on $\mathbf{X}$, the target of estimation is $\boldsymbol{\theta}(\boldsymbol{P})$, whereas $\boldsymbol{\theta}(\mathbf{X})$ is a random quantity (see Corollary 4.3 below).

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

- Functionals defined through minimization of objective functions:

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

- Functionals defined through estimating equations:

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

These specialize to normal equations for linear OLS by (4).

4.3 Estimation Offsets

In Part I we defined what we call "estimation offsets." With the availability of $\boldsymbol{\theta}(\mathbf{X})$ for arbitrary regression functionals, these can be defined in full generality:

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

The total EO is the deviation 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 component due to the randomness of $\vec{\boldsymbol{X}}$ in the presence of misspecification. These facts will be elaborated in what follows.

The approximation EO lends itself for another characterization of well-specification in the sense of Section 3:

Proposition 4.3: *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 $\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}(\boldsymbol{P}) = \mathbf{0}$ for all acceptable $\mathbf{X}$.*

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

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

The converse follows because the empirical regressor distributions $\hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ (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 $\mathbf{X}$ 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 $\mathbf{X}$ is random:

Corollary 4.3: *Same assumptions as in Proposition 4.3.*

- *Fixed-$\mathbf{X}$ and random-$\mathbf{X}$ 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 $\boldsymbol{V}[\boldsymbol{\theta}(\mathbf{X})] = \mathbf{0}$ 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.

4.4 Deterministic Association Annihilates the Noise EO

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

Proposition 4.4: *If* $Y = f(\vec{\boldsymbol{X}})$ *is a deterministic function of* $\vec{\boldsymbol{X}}$*, then* $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\mathbf{X}) \overset{\boldsymbol{P}}{=} 0$ *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 $\hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}$ is $\hat{\boldsymbol{P}}$: $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}} = \hat{\boldsymbol{P}}$. It follows that $\boldsymbol{\theta}(\mathbf{X}) = \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}}) = \boldsymbol{\theta}(\hat{\boldsymbol{P}}) = \hat{\boldsymbol{\theta}}$. □

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 the case of linear OLS, the conditional response distribution $Y|\vec{\boldsymbol{X}}$ is the driver of the noise EO.

In Part I, Section 6, we used deterministic responses as illustrations where there is no noise EO and the only sampling variability in $\hat{\boldsymbol{\theta}}$ stems from the randomness of the regressors interacting with misspecification — the second of the two “conspiracy” effects of the title of Part I.

4.5 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 one can make is in terms of influence functions; for linear OLS it translates to this correspondence. Using again the notations $\boldsymbol{IF}(y, \vec{\boldsymbol{x}})$ and $\boldsymbol{E}[\boldsymbol{IF}(Y, \vec{\boldsymbol{X}})|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \boldsymbol{IF}(\vec{\boldsymbol{x}})$ for the full and the partial influence functions of $\boldsymbol{\theta}(\cdot)$ (Section 3.4), and assuming asymptotic linearity of $\boldsymbol{\theta}(\cdot)$, the EOs have the following approximations to order $o_P(1/\sqrt{N})$:

$$\boxed{\begin{array}{lll} \textit{Total EO:} & \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{P}) & \approx \frac{1}{N}\sum_i \boldsymbol{IF}(Y_i, \vec{\boldsymbol{X}}_i), \\ \textit{Noise EO:} & \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\mathbf{X}) & \approx \frac{1}{N}\sum_i \Big(\boldsymbol{IF}(Y_i, \vec{\boldsymbol{X}}_i) - \boldsymbol{E}[\boldsymbol{IF}(Y, \vec{\boldsymbol{X}}_i)|\mathbf{X}_i]\Big), \\ \textit{Approx. EO:} & \boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}(\boldsymbol{P}) & \approx \frac{1}{N}\sum_i \boldsymbol{E}[\boldsymbol{IF}(Y, \vec{\boldsymbol{X}}_i)|\mathbf{X}_i]. \end{array}} \tag{8}$$

The first and the third approximation are standard under the assumption of asymptotic linearity because $\boldsymbol{IF}(y, \vec{\boldsymbol{x}})$ is the influence function for $\boldsymbol{\theta}(\boldsymbol{P})$, while $\boldsymbol{IF}(\vec{\boldsymbol{x}}) = \boldsymbol{E}[\boldsymbol{IF}(Y, \vec{\boldsymbol{X}})|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}]$ is the influence function of $\boldsymbol{\theta}(\mathbf{X})$. The second approximation is the difference of the first and the third. The approximations to the noise EO and the approximation EO are orthogonal to each other, hence the two EOs are asymptotically uncorrelated. — These approximations lead straight to the CLTs of the next section.

5. MODEL-ROBUST CENTRAL LIMIT THEOREMS DECOMPOSED

The purpose of the following CLTs is to decompose the asymptotic sampling variability of plug-in estimates into two parts:

- a contribution due to the conditional variability of the response, and
- a contribution due to the marginal variability of the regressors,

the latter being the "conspiracy contribution" to sampling variability caused by regressor randomness interacting with misspecification in the sense of Section 3.

5.1 CLT Decompositions Based on Influence Functions

As in Section 4.5 consider a regression functional $\boldsymbol{\theta}(\boldsymbol{P})$ that is asymptotically linear with influence function $\boldsymbol{IF}(y,\vec{\boldsymbol{x}})$ and partial influence function $\boldsymbol{IF}(\vec{\boldsymbol{x}}) = \boldsymbol{E}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}\!=\!\vec{\boldsymbol{x}}]$. The EOs obey the following CLTs:

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

These are immediate consequences of the assumed asymptotic linearities of Section 4.5. The asymptotic variances of the EOs follow the canonical decomposition

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

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 $\boldsymbol{IF}(Y,\vec{\boldsymbol{X}}) - \boldsymbol{E}[\boldsymbol{IF}(Y,|\vec{\boldsymbol{X}})$ and $\boldsymbol{E}[\boldsymbol{IF}(Y,|\vec{\boldsymbol{X}})$ are orthogonal to each other, which implies by (8) that the noise EO and the approximation EO are asymptotically orthogonal. For linear OLS this orthogonality holds exactly for finite N due to (6) and (12) in Part I: $\boldsymbol{V}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\mathbf{X}), \boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P})] = \mathbf{0}$.

The following corollary is a restatement of Proposition 3.4, but enlightened by the fact that it relies on the asymptotic variance of the approximation EO.

Corollary 5: *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 $\boldsymbol{V}_{\boldsymbol{P}_{\vec{\boldsymbol{X}}}}[\boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})|\vec{\boldsymbol{X}}]] = \mathbf{0}$ for all acceptable $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. This in turn means $\boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}}[\boldsymbol{IF}(Y,\vec{\boldsymbol{X}})|\vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}] = \mathbf{0}$ for all $\vec{\boldsymbol{x}}$, which is the condition of Proposition 3.4. □

5.2 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 $\boldsymbol{\Lambda}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})]$ is the Jacobian of size $q \times q$, $q = \dim(\boldsymbol{\psi}) = \dim(\boldsymbol{\theta})$. Then the CLTs specialize to the following:

$$\boxed{\begin{array}{rcl} \sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) & \xrightarrow{\mathcal{D}} & \mathcal{N}\Big(\mathbf{0},\, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{V}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})]\,\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{\,-1}\Big) \\ \sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\mathbf{X})) & \xrightarrow{\mathcal{D}} & \mathcal{N}\Big(\mathbf{0},\, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{E}[\boldsymbol{V}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,]\,\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{\,-1}\Big) \\ \sqrt{N}\,(\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta}) & \xrightarrow{\mathcal{D}} & \mathcal{N}\Big(\mathbf{0},\, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{V}[\boldsymbol{E}[\boldsymbol{\psi}(\boldsymbol{\theta};Y,\vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,]\,\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{\,-1}\Big) \end{array}}$$

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

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

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 $\boldsymbol{IF}(y, \vec{\boldsymbol{x}}) = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}(\vec{\boldsymbol{x}}\vec{\boldsymbol{x}}'\boldsymbol{\beta} - \vec{\boldsymbol{x}}y)$, $\boldsymbol{\psi}(\boldsymbol{\beta}; y, \vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}}\vec{\boldsymbol{x}}'\boldsymbol{\beta} - \vec{\boldsymbol{x}}y$ and $\boldsymbol{\Lambda} = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']$. Hence the CLTs of Part I, Proposition 7.

5.3 Implications of the CLT Decompositions

The most impactful part of the CLT decompositions concerns the approximation EO because it may be subject to confusions: Misspecification, in traditional parametric modeling, is sometimes called "model bias" which, due to unfortunate terminology, may suggest a connection to estimation bias, $\boldsymbol{E}[\hat{\boldsymbol{\theta}}_N] - \boldsymbol{\theta}(\boldsymbol{P})$. Importantly, there is no connection between the two notions of bias. Estimation bias vanishes at a rate faster than $1/\sqrt{N}$ and does not contribute to standard errors derived from asymptotic variances. Model bias, on the other hand, which is misspecification, generates a contribution to the standard error, and this contribution is asymptotically of order $1/\sqrt{N}$, the same order as the better known contribution due to the conditional noise in the response. This is what the CLT decompositions reveal. 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.*

To make sense of this statement it is necessary to assume the point of view that all quantities of interest are regression functionals: Their plug-in estimates have two sources of sampling variability, both of the same order, the better known source being $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and the lesser known source being $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ in conjunction with model bias/misspecification.

6. PLUG-IN/SANDWICH ESTIMATORS VERSUS M-OF-N BOOTSTRAP ESTIMATORS OF STANDARD ERROR

6.1 Plug-In Estimators are Limits of M-of-N Bootstrap Estimators

In Part I, Section 8, it was shown 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: Because resampling is iid sampling from some distribution, there holds a CLT as the resample size grows, $M \to \infty$. The distribution being (re)sampled is the empirical distribution $\hat{\boldsymbol{P}}_N = (1/N)\sum \delta_{(y_i, \vec{\boldsymbol{x}}_i)}$, where N is fixed but $M \to \infty$ (causing ever more duplicates of cases). The following holds for bootstrap resampling of any well-behaved statistical functional, be it in a regression context or not:

Proposition 6: *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:*

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

Let a fixed dataset of size N *with acceptable regressors be represented by the empirical measure* $\hat{\boldsymbol{P}}_N$*. 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* $\hat{\boldsymbol{P}}_N$*, we have:*

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

The proposition contains its own proof.

Corollary 6: *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*.*

6.2 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 subsection. The result states that so-called "M-bagged functionals" are "smooth" in a certain sense, the smoother the smaller the resample size M. In this sense, the limit $M \to \infty$ is the least smooth choice that is more likely to behave erratically than all choices of finite M. We next define "bagged functionals", explain the meaning of "smoothness", and describe the link of this result to the issue of "bootstrap versus plug-in/sandwich estimators".

"Bagging" is "bootstrap averaging" as introduced by Breiman (1996). While bootstrap and bagging are applied in practice to empirical distributions $\hat{\boldsymbol{P}}_N$, 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

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

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 any order, where the first order term is the average of the first order influence function (Section 4.5) that drives asymptotic normality (Section 5). Higher order terms reflect "interactions" similar to U-statistics (actually, V-statistics) based on pairs, triples, ... of observations. It is intuitively clear that an M-bagged functional cannot have interactions of order higher than M. This is 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 smoothness for bagged functionals that is analogous to the natural notion of smoothness of polynomials: A polynomial is the smoother the lower its degree; similary, a bagged functional is the smoother the lower its bag size M. In both cases it is the length (or better: "shortness") of the expansion that is a measure of "smoothness."

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 trivially as follows:

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

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

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

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:

$$\boldsymbol{P} \mapsto \boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)] \quad\text{and}\quad \boldsymbol{P} \mapsto \boldsymbol{AV}[\boldsymbol{P};\boldsymbol{\theta}(\cdot)]. \tag{9}$$

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

$$\boldsymbol{BV}_M[\boldsymbol{P};\boldsymbol{\theta}(\cdot)] \;=\; M\cdot\big(\,\boldsymbol{E_P}[\boldsymbol{\theta}(\boldsymbol{P}_M)^2] - \boldsymbol{E_P}[\boldsymbol{\theta}(\boldsymbol{P}_M)]^2\,\big)\,,$$

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

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

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

The practical consequence of the preceding argument is that, in terms of sampling variability, bootstrap standard errors are likely to be more stable than plug-in/sandwich estimators of standard error. Most likely the greatest stability gains are obtained for very 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.

Whether this stability mitigates the problem of outlier/heavy-tail non-robustness of sandwich estimators noted in Part I, Section 13, we would not know. As noted earlier, neither do we have specific recommendations for the choice of the resample size M in a given data analytic situation.

7. 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 mis- and 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. These findings are the two "conspiracy effects" whereby nonlinearity and randomness of the regressors combine ("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 independent of the conditional randomness of 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? A step toward an answer is to leverage the basic structure that defines a regression problem: the asymmetric analysis of the association between Y and $\vec{\boldsymbol{X}}$ in terms of the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. This suggests taking the joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ apart and analyze the issue of mis- and well-specification in terms of $\boldsymbol{\theta}(\cdot)$, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. The solution, finally, to

- establishing a compelling notion of mis- and well-specification at this level of generality, and
- extending the two "conspiracy effects" to arbitrary types of regression,

is to look no further and use the first conspiracy effect as the definition of misspecification: dependence of the regression functional on the regressor distribution. The second effect is then a corollary of the definition: If the target $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \boldsymbol{P}_{\vec{\boldsymbol{X}}})$ is non-constant in $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, so is the conditional target $\boldsymbol{\theta}(\mathbf{X}) = \boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \otimes \hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}})$, which hence is random and contributes to the overall sampling variability of the estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{P}})$.

Intuitively this notion of misspecification means that whatever is measured by $\boldsymbol{\theta}(\cdot)$ depends on where in regressor space the observations have fallen. 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.

It is a peculiar fact that under misspecification the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ moves center stage together with $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. This is unexpected because regression, as traditionally understood, attempts to describe aspects of the conditional response distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ while essentially sidelining the regressor distribution by conditioning on the observed regressors. The justification for conditioning is the ancillarity argument applied to the regressors, but as was shown in Part I this argument no longer applies under misspecification. Regressor ancillarity has a curious mirror image in the present theory of regression functionals: Well-specification means that for the conditional response distribution at hand, the regression functional is not really a function of the regressor distribution, thereby providing a justification for sidelining it. However, in a realistic view of regression analysis, well-specification is a mere ideal, while degrees of misspecification are the reality. There exists therefore little justification for conditioning on the observed regressors and for treating them as fixed because implicit is the assumption of well-specification.

With degrees of misspecification accepted as the reality in regression analysis, we of course do *not* advocate carelessness in its practice. It should always be mandatory to perform due diligence in terms of regression diagnostics, and in fact we proposed new types of diagnostics in Section 3.5. There is, however, an argument to be made to feel less guilty about using simple models when they capture essential features of the data. In addition, there is an argument to be made in favor of using statistical inference that is model-robust, and to this end one can use either x-y bootstrap estimators or plug-in/sandwich estimators of standard error. Between the two we advanced some arguments in favor of using bootstrap over plug-in/sandwich estimators. More importantly, however, both approaches to inference are in accord with the following important principle: When regressors are random, they should be treated as random.

Acknowledgments: We are grateful to Gemma Moran and Bikram Karmakar for their help in the generalizations of Section 4.2.

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

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

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

[14] 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.

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

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

[17] 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.

[18] 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.

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

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

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

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

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

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

APPENDIX

A. Proper Scoring Rules, Bregman Divergences and Entropies

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.

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 $\boldsymbol{E_P}[\cdot]$, 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 $\boldsymbol{E_P}[\cdot]$, 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 $D_{KL}(\boldsymbol{P},\boldsymbol{Q}) = \boldsymbol{E_P}[-\log q(Y)] - \boldsymbol{E_P}[-\log p(Y)]$, 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) $D(\boldsymbol{P},\boldsymbol{Q}) \;=\; \boldsymbol{E_P}[S(Y,\boldsymbol{Q})] - H(\boldsymbol{P})$.

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 $S(\boldsymbol{P},\boldsymbol{Q}) := \boldsymbol{E_P}[S(Y,\boldsymbol{Q})]$. 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. For insightful background on proper scoring rules, see Gneiting and Raftery (2007) (but note two reversals of conventions: their $S(\boldsymbol{Q},\boldsymbol{P})$ corresponds to our $-S(\boldsymbol{P},\boldsymbol{Q})$).

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 α 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.

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

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

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:

$$\boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}}[\mathcal{L}(\boldsymbol{\theta}_0; Y, \vec{\boldsymbol{x}})] \leq \boldsymbol{E}_{\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}}[\mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{x}})] \quad \forall\boldsymbol{\theta},\ \forall\vec{\boldsymbol{x}}. \tag{10}$$

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

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 $\phi(q)\!=\!q^2$.

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 familiy 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:

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

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:

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

B. 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 3.4:

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

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 3.4. 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 $\boldsymbol{P}$:

$$(11)\quad \frac{d}{dt}\Big|_{t=0}\,\boldsymbol{\theta}(\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}\otimes((1-t)\boldsymbol{P}_{\vec{\boldsymbol{X}}}{}'+t\boldsymbol{P}_{\vec{\boldsymbol{X}}})) \;=\; \int\!\!\int \boldsymbol{IF}(y,\vec{\boldsymbol{x}})\,\boldsymbol{P}(dy|d\vec{\boldsymbol{x}})\boldsymbol{P}'(d\vec{\boldsymbol{x}}),$$

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 3.4, 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 $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. □

C. Miscellaneous Proofs

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 $\tilde{\boldsymbol{P}}_{\boldsymbol{X}}=(1\text{–}t)\boldsymbol{P}_{\vec{\boldsymbol{X}}}+t\delta_{\vec{\boldsymbol{x}}}$ for $0<t<1$ will then also be acceptable (see Section 2.1), 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. □