Statistical Science
2019, Vol. 34, No. 4, 606–620
https://doi.org/10.1214/19-STS762
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Models as Approximations—Rejoinder

Andreas Buja, Arun Kumar Kuchibhotla, Richard Berk, Edward George, Eric Tchetgen Tchetgen and Linda Zhao

Abstract. We respond to the discussants of our articles emphasizing the importance of inference under misspecification in the context of the reproducibility/replicability crisis. Along the way, we discuss the roles of diagnostics and model building in regression as well as connections between our well-specification framework and semiparametric theory.

Key words and phrases: Well-specification, reproducibility/replicability, proper scoring rules, causal inference, semiparametrics, diagnostics.

"Truth is much too complicated to allow anything but approximations." (John von Neumann, as cited by Aronov and Miller, 2019*)*

We are grateful to the Editors of *Statistical Science* for publishing our two articles together, referred to herein as "Part I" and "Part II." In combination, they represent themes in recent work by a group of Wharton faculty and students centered around the late Larry Brown. Based on this work, Larry initiated follow-up efforts that extended what we had learned to the two-stage bootstrap (McCarthy et al., 2018), to semisupervised learning (Azriel et al., 2016), and to the estimation of population Average Treatment Effects (ATEs) (Pitkin et al., 2013).

We also thank the discussants for investing time and effort to comment on our work. We are fortunate to have the opportunity to reargue and qualify aspects of misspecified models and quantities of interest in light of their comments. The idea of models as approximations is now finding explicit treatment in excellent books by Davies (2014) and Aronov and Miller (2019).

Andreas Buja is the Liem Sioe Liong/First Pacific Company Professor of Statistics. Arun Kumar Kuchibhotla is Doctoral Student of Statistics. Richard Berk is Professor of Criminology and Statistics. Edward George is the Universal Furniture Professor of Statistics. Eric Tchetgen Tchetgen is the Luddy Family President's Distinguished Professor. Linda Zhao is Professor of Statistics. Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: lzhao@wharton.upenn.edu).

It may be worthwhile to begin by highlighting the most important ideas of our articles. They are found in Part II, Sections 3–5. These ideas include the notion of well-specification of quantities of interest (as opposed to correct specification of models) and diagnostics for well-specification based on reweighting. Instructive ideas are often simple, as is this: *the purpose of data analysis is to extract meaningful quantities*, not models. By emphasizing quantities of interest rather than models, we step beyond "Models as Approximations" in our titles. Quantities of interest are most frequently model-based, but they can be constructed in other ways as well, including ad hoc constructions (as in Part II, Section 3.2, item 3) or based on subject matter expertise. The next idea also is simple and fundamental: *regression is the attempt to characterize the conditional response distribution.* This implies an attempt to find properties of the data distribution that do not depend on the regressor distribution. It is then a small step to *define well-specification as the irrelevance of the regressor distribution for the quantities of interest.* Consequently, well-specified quantities characterize the conditional response distribution alone. The final fundamental idea is to diagnose well-specification of quantities of interest by applying them to *reweighted data, where the weights depend on the regressors only.* The justification follows from the fact that regressor-dependent reweighting changes the regressor distribution but not the conditional response distribution. To move from motivating principles to concrete illustrations, we suggest that readers turn to the data-driven example shown in Figure 1 of Part II for a demonstration of the well-specification diagnostic:

Not only does reweighting detect misspecification of a quantity of interest, it also pinpoints its nature without further modeling.

Turning to the discussants' comments, it is a pleasure to learn from their thoughtful comments, both technical and methodological. Some reveal strong differences in views:

- On the one hand, statisticians such as Jerald Lawless assume that model-robust sandwich standard errors and the consequences of misspecification are widely known and accepted. They tend to be informed by theories of estimating equations and econometrics.
- On the other hand, statisticians such as Antony Davison and coauthors deeply object to undoing long established model-trusting practice. They tend to be informed by diagnostic methodologies that may proffer evidence of model (in)adequacy.—A second school of statisticians who are inclined to remain model-trusting includes Bayesians such as Rod Little.

In the substance, we strongly agree with the first position. Degrees of misspecification are the old and the new normal; models have never been more than approximations. Misspecification, especially in rich models, may not be detectable, but even if it is, we may have reasons not to act on it: (1) The type of misspecification may not matter for the purpose at hand; (2) our subject matter clients may require a misspecified model for simplicity, or for comparability to extant literature, or because of subject matter theories; (3) large data problems with many regressors and many responses may require fitting large numbers of models for which detailed diagnostics may be unrealistic and simplicity may be a higher priority; (4) a data analysis plan may have been preregistered or otherwise specified a priori, but in retrospect was based on a misspecified model. For these reasons and others, misspecification may need to be tolerated, even if detectable with diagnostics, and *assumption-lean* inference may be needed.

We draw special attention to the last of these reasons, (4) preregistration of data analysis plans (e.g., Adam, 2019) because it may play an increasingly important role in the critical context of the reproducibility/replicability crisis (e.g., Ioannidis, 2005).[1] Pre-registration has been proposed as a way to reduce the rate of false empirical findings caused by informal and/or unreported analysis steps. It has been recognized that certain data analytic practices that used to be acceptable can invalidate statistical inference (e.g., Simmons, Nelson and Simonsohn, 2011). Jerald Lawless hits the nail on the head when he writes that "Exploratory data analysis and alternating bouts of model fitting and assessment are used in many settings, and create difficulties for formal inferences concerning covariate effects and 'final' models." For example, a powerful component of EDA and diagnostics is data visualization, but data analytic decisions based on them tend to be informal.[2] Another form of data exploration is trial and error experimentation where regressors and responses are selected by manually trying out multiple models. At a higher level, trial and error experimentation also takes place when multiple selection algorithms[3] are applied and their results compared and "meta-selected" based on informal preferences. Further experimentations may occur with basis expansions, addition of interaction terms, variable transformations, outlier removal and subsetting.

To appreciate the problems created by informal practices, one must exercise the frequentist imagination by conjuring a multitude of alternative datasets drawn from the same population and the ensuing variability of data analytic results. Even when all is done in the name of careful and competent data analysis, with the hope of finding correctly specified models justifying model-trusting inference, the results may be too good to be true. The tea leaves might have been arranged too carefully. Competent data analysis may inadvertently turn into data dredging, contributing to the reproducibility/replicability crisis.

As a partial countermeasure at the methodological level, efforts have been launched by some journals and funding agencies to require or reward preregistration of data analysis procedures before examining

[1]The use of the terms *reproducibility* and *replicability* is not consistent in the literature. We frame our intended meaning in an idealized frequentist way as follows: When multiple isomorphic datasets (e.g., same variables, same number of cases) are obtained from the same or related data sources (experiments, surveys, clinical trials, ...), identical statistical analyses applied to them produce results that agree with each other to a degree that is consistent with their nominal error probabilities. The terms *isomorphic, identical, same* and *related* pose semantic issues in need of clarification and consensus. Practical issues arising from multiple datasets with related data sources are the domain of meta-analysis.

[2]Examples: Does a scatterplot suggest a variable transformation? Is a plot of residuals versus fitted values so flawed that the model should be modified?

[3]Examples include lasso, stepwise, all subsets, combined with various criteria for model size.

the data.[4] Consequently, we will be confronted by various degrees of misspecification and the need to introduce assumption-lean inferences, even if misspecification is detectable after the fact. Nothing prevents researchers from going beyond planned or preregistered procedures, but such steps must be clearly identified as exploratory rather than confirmatory. Likewise, our newly proposed diagnostics belong more in the exploratory than confirmatory realm (unless preregistered as part of a larger data analysis plan), and this is despite inferential features such as null distributions for the ***RAV*** test (Part I, Sections 12.2–3) and bootstrap bands for the reweighting diagnostic (Part II, Sections 4–5). Against Davison and coauthors' comfort with diagnostics-driven model building and subsequent model-trusting inferences, learning from data demands a trade-off: *if confirmation is required, it limits exploration, and if exploration is required, it limits confirmation.*

Having framed our theory of mis/well-specified regression in the larger context of reproducibility/replicability, we proceed to specific comments made by the discussants.

Jerald Lawless provides a sympathetic interpretation of our articles, and he also makes valuable observations from a more holistic point of view. We agree with him that some communities accept model-robustness and sandwich estimators, but we can't quite agree that the consequences of misspecification are widely understood. Part I is about explaining these consequences in detail and correcting misconceptions. The most prevalent among them derives from using the term "model bias" as a synonym for misspecification, which suggests erroneously that model bias creates estimation bias. As explained at the end of Section 5 of Part I and again in Section 7.3 of Part II, what really happens is that model bias (misspecification) funnels the randomness of the regressors into sampling variability in estimates, thereby contributing to their standard errors rather than their bias (Figure 4, Part I). This is why regressor randomness matters, why treating regressors as fixed is often wrong, and why their randomness should be accounted for by statistical inference. Assumption-lean standard errors get this right.

[4] An example of impressive discipline is reported from the physicists at CERN in the context of the Higgs Boson discovery: "Once we look at the real data, ... we're not allowed to change the analysis anymore" (Hartman, 2014). An important post hoc requirement is disclosure of all analysis steps taken, not just those that led to "significant" results.

Except when they don't. We note in Part I, Section 13, that sandwich standard errors for linear OLS are nonrobust to outliers and heavy tails, as is apparent from equation (24) in Section 12.1 of Part I. In addition, classical heavy-tail robustness and model robustness are in conflict at the level of standard error estimates. Among possible approaches are classically robust methods (see, e.g., Cantoni and Ronchetti, 2001) and, if meaningful, transformations to bounded ranges or at least well-behaved tails (Part I, Section 13), preferably chosen a priori to avoid the data dredging problem.

Jerald Lawless mentions the distinction, which should be common, between scientific discovery, explanation and understanding on the one hand, and automatic decision making and prediction on the other hand. Our articles address the former, and it is here where statistical inference for parameters/regression functionals plays a critical role. Prediction, in contrast, typically relies on some form of cross-validation. Then again, one may also consider prediction-related functionals, such as $\boldsymbol{\beta}(\boldsymbol{P})'\vec{\boldsymbol{x}}_0$ obtained from linear OLS regression, that is, the model-predicted approximation to the conditional response mean $\mu(\vec{\boldsymbol{x}}_0)$ at $\vec{\boldsymbol{x}}_0$. Model-robust standard error estimates can be obtained from the sandwich formula or the x-y bootstrap. Well-specification of the prediction functional can be examined with the reweighting diagnostic. Significant nonconstancy under reweighting may indicate that the true conditional mean function $\mu(\cdot)$ is highly nonlinear or that the location $\vec{\boldsymbol{x}}_0$ is an extrapolation. If the weight functions are centered at $\boldsymbol{\xi} = \vec{\boldsymbol{x}}_0$ with shrinking bandwidth, one obtains locally linear fits that amount to nonparametric function estimation (Part II, Appendix A.7). A reweighting approach applies if interest is solely in point predictions. More often, however, interest is in prediction intervals for a quantitative response, and to this end one can use empirically calibrated prediction bands (Berk et al., 2019, Section 9). These are entirely model-robust, though not optimized for any particular location $\vec{\boldsymbol{x}}_0$. Modern theories of prediction have been proposed by Lei et al. (2018) and Steinberger and Leeb (2018).

We agree and have pointed out (Part I, Section 6.2) that model assessment is difficult when the number of covariates is greater than one. To fully appreciate the effects shown in the figures of Part I for $p = 1$, we have to tax our imagination and conceive of the analogs for $p > 1$ and even the "modern" case, $p > n$.

Finally, Jerald Lawless wonders whether our insights could also be found with standard results based on

working models under model misspecification. This is indeed done in Part II, Section 7.2, where we show the usual CLT for functionals defined by estimating equations (EE), a special case being the score equations of a working model. In addition, we point out a Pythagorean decomposition of the CLT into a well-known part due to the conditional response distribution $\boldsymbol{P}_{Y|\vec{X}}$ and a lesser known part due to the regressor distribution $\boldsymbol{P}_{\vec{X}}$, mediated by misspecification of the EE functional. This second part vanishes under well-specification of the EE functional. An EE functional is well-specified if there exists a parameter setting for which the estimating equations are satisfied conditionally at (a.s.) all locations in regressor space (Proposition 3.3 of Part II). A misspecified EE functional satisfies the equations only on average wrt $\boldsymbol{P}_{\vec{X}}$, not pointwise.

Sara van de Geer begins with a nice formulation of the functional view of parametric statistics: "... view a parameter as some function of the distribution instead of thinking of the distribution as some function of a parameter." She then proceeds to a reconstruction of Huber's (1967) sandwich result for general M-estimators, all in 1.5 pages. At the end of her Section 1, she shows for M-estimators how the meat of the sandwich covariance matrix consists of the inverted bread plus a Hessian ingredient caused by misspecification. Collapse of the sandwich to merely one slice of bread occurs when the Hessian ingredient vanishes. Such collapse is assumed by the practitioners of model-trusting inference (statistical vegetarians of sorts, although food analogies suffer at this point).

While Sara van de Geer's Section 1 is about M-estimation in general, her Section 2 turns to M-estimation for regression. Of particular interest is a generalization from our linear OLS context in Part I to generalized linear regressions obtained from M-estimators that minimize loss functions of the form $\rho(\vec{x}'\boldsymbol{\beta}, y)$. In this broad context, she shows how it is possible to generalize the notions of conditional response surfaces $\mu(\vec{x})$, conditional variances $\sigma^2(\vec{x})$ and nonlinearities $\eta(\vec{x})$. The conditional MSE decomposition $m^2(\vec{x}) = \sigma^2(\vec{x}) + \eta^2(\vec{x})$ of our equation (8) in Part I generalizes immediately, as does the ensuing decomposition of the CLT and its asymptotic sandwich variance of our Proposition 7.1 in Part I.

Sara van de Geer also elaborates on the classical view of ancillarity, reverting to an interpretation of distributions as functions of parameters. This is helpful because in our experience, many statisticians do not know that regressor ancillarity is the classical argument to justify the treatment of regressors as fixed when they are truly random (assumed here throughout). We recommend following her explanation of nonancillarity in the classical sense. The regression parameters $\boldsymbol{\beta}$ would also be parameters (in the classical sense) of a model for the regressor distribution: $-\log p_{\boldsymbol{\beta}}(\vec{x}, y) = -\log p_{\boldsymbol{\beta}}(y|\vec{x}) - \log p_{\boldsymbol{\beta}}(\vec{x})$. Hence the *regressor model* $p_{\boldsymbol{\beta}}(\vec{x})$ would compete with the *regression model* $p_{\boldsymbol{\beta}}(y|\vec{x})$ to determine the (classical) parameter $\boldsymbol{\beta}$.

Finally, we agree that the move from classical parameters to regression functionals requires rethinking of meanings. We prefer, however, to modify Sara van de Geer's last sentence as follows: If a regression functional derives from an approximating regression model, state the meaning of the estimates as usual in a model-trusting manner, but add the clause "... in the best approximating model."

Dag Tjøstheim also provides a sympathetic assessment of our perspective. We appreciate his remark on our proposal for interpreting the meaning of linear OLS slopes in the presence of misspecification (Part I, Section 10). Although the interpretation seems straightforward, we readily acknowledge that its generalizability beyond linear OLS is limited. We maintain, however, that misspecification is often undetectable, more so with increasing model complexity, and even if detectable, misspecification may be imposed for reasons discussed earlier. We agree with Dag Tjøstheim that Part II is more satisfying than Part I, although necessarily more opaque in the treatment of asymptotics. Most satisfying to us are, as mentioned earlier, Part II, Sections 3–5.

Dag Tjøstheim is correct in thinking that we are not parametric extremists, and that we consider nonparametric methods fair game. We are simply responding to the reality that parametric methods still dominate current practice. Moreover, the reweighting diagnostic of Part II has an advantage over nonparametric fitting in that it allows analysts to discover the nature of misspecification and yet remain within their misspecified model. Dag Tjøstheim notes, as we do in Appendix A.7 of Part II, that Parzen kernel methods generate localized functionals such as $a(\vec{x}_0) = \boldsymbol{E}[YK(\vec{X} - \vec{x}_0)]/\boldsymbol{E}[K(\vec{X} - \vec{x}_0)]$, which is a reweighted response mean. Conventional Parzen kernel methods will generally be useful in very low regressor dimensions such as $p = 1$ or 2. A difference with our reweighting diagnostic is that we consider reweighting kernels that depend on a single regressor or other

one-dimensional quantity derived from, or related to, regressors. This kind of localization is less vulnerable to the curse of dimensionality. It also is highly interpretable because it is based on the original model for arbitrary quantities of interest and lends itself to plotting, as illustrated by Figures 1–3 of Part II. We mention peripherally that there exists a limited connection to one-term additive modeling (Part II, Section 5.2 and Appendix A.6), but the reweighting diagnostic is capable of detecting interactions as well, as illustrated by Figure 1 of Part II.

Dag Tjøstheim makes some interesting forays into the realm of dependent data such as time series. Here, the assumption of i.i.d. data made in our articles is of the "assumption-rich" variety. Our methods do not generalize, for example, to time series autoregressive models. For one thing, the regressors are lagged responses and cannot be used for regressor-dependent reweighting. Reweighting according to the time dimension is a possibility, amounting to time-localization of the model. Well-specification then amounts to stationarity of the quantity of interest. On different grounds, some of us (Kuchibhotla et al., 2018a, 2018b) have produced work that justifies model-robust inference under certain types of dependence.

Rod Little provides an interesting foil through his Bayesian perspective. As expected, we fail to convince each other, but we have an important point of agreement: Bayesian approaches are fundamentally model-trusting as stated in the Conclusions of Part I. At the root is the Bayesian idea that the parameters are uncertain/random, whereas the data are certain/fixed. As a consequence, "conditioning on the data" (not just the regressors) conceals that the data could have been different and that the likelihood as a probabilistic model of datasets could have been misspecified. Bayesians who wish to calibrate their inferences in a frequentist sense, as does Rod Little, must find ways to match the frequentist variability of estimates across datasets with the variability of the parameters drawn from the posterior, at least approximately in terms of first- and second-order moments, to equalize asymptotic normal distributions.

A further point of agreement with Rod Little is that models as approximations to the truth can be useful. He recommends the Bayesian reconstruction of sandwich-based inference by Szpiro, Rice and Lumley (2010), an article we cited but did not discuss in detail, offering us an opportunity to do so now. This is indeed a heroic exercise in calibrated Bayesian inference to asymptotically match frequentist model-robust inference in linear regression. Here is their logic.

The article starts out by defining a target of inference in their equation (3) using notation that differs from ours in Section 3.2, Part I. This cosmetic difference should not obscure that we define the same target: the best linear approximation to the response Y or, equivalently, to the conditional response mean $\mu(\vec{x}) = \boldsymbol{E}[Y|\vec{\boldsymbol{X}} = \vec{x}]$. The coefficients of this linear approximation are in our notation

$$\boldsymbol{\beta}(\boldsymbol{P}) = \operatorname*{argmin}_{\boldsymbol{\beta}} \boldsymbol{E}_{\vec{\boldsymbol{X}}}\big[\big(\mu(\vec{\boldsymbol{X}}) - \vec{\boldsymbol{X}}'\boldsymbol{\beta}\big)^2\big]. \tag{1}$$

Szpiro et al. continue with a Bayesian construction that puts priors on the conditional mean function $\mu(\vec{x})$, the conditional variance function $\sigma^2(\vec{x}) = \boldsymbol{V}[Y|\vec{\boldsymbol{X}} = \vec{x}]$, and the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. Assuming a normal conditional likelihood $Y|\vec{\boldsymbol{X}} \sim \mathcal{N}(\mu(\vec{\boldsymbol{X}}), \sigma^2(\vec{\boldsymbol{X}}))$, they obtain posteriors for $\mu(\cdot)$, $\sigma^2(\cdot)$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. From these, they derive a posterior distribution for the coefficients $\boldsymbol{\beta}$ of the best linear approximation by applying formula (1) above to the draws of $\mu(\cdot)$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ from the posterior. For discrete regressor distributions, $\boldsymbol{P}_{\vec{\boldsymbol{X}}} = \sum_{k=1,\ldots,K} \lambda_k \delta_{\boldsymbol{\xi}_k}$, they show (ibid., Theorem 1) that the posterior expectations of the coefficients approximate the OLS estimates, and the posterior standard deviations approximate the sandwich standard errors for increasing sample sizes. This is not problematic for discrete regressor distributions because at each discrete regressor location $\boldsymbol{\xi}_k$, one will ultimately see multiple observations. It comes down to estimating means of Y at each $\vec{\boldsymbol{x}} = \boldsymbol{\xi}_k$, but even this simple case requires a multipage proof rich in "mathematicity" (online suppl., Szpiro et al., 2010). Generalization to continuous regressor distributions requires nonparametric Bayesian function estimation in p dimensions for $\mu(\cdot)$ and $\sigma^2(\cdot)$. This is carried out by Szpiro et al. in practical terms for a single regressor, $p = 1$, based on regression splines.

In a cultural disconnect, Rod Little sees the Szpiro et al. construction as fundamentally simple and even enjoyable, whereas our frequentist analysis is in his mind overly mathematized and difficult to explain to practitioners. From a frequentist perspective, things may look different; it is an additional burden to explain the complexities in constructing priors, which drop out asymptotically to approximate in the limit the OLS estimates and their sandwich-based standard errors. In the case of nondiscrete regressor distributions, another

burden is estimating the infinite dimensional parameters $\mu(\cdot)$ and $\sigma^2(\cdot)$ for inference about the finite dimensional parameters of a linear approximation.

The frequentist analysis of Part I and II is free of such complexities and to the point, which is that the true assumption-lean sampling variability of estimates is composed of two sources, one due to $Y|\vec{\boldsymbol{X}}$ and one due to $\vec{\boldsymbol{X}}$ in the presence of misspecification. As Peter Aronov (personal communication) put it, simply and concisely: "Under misspecification, design matters." We see few signs that this point is appreciated, even though we emphasize it in the finite sample analysis (Part I, Sections 5–6), in the CLTs (Part I, Section 7), and again in Part II (Sections 6–7).[5]

We agree with Rod Little that our writing is rich in mathy notation and formulas. The intent was precision of thinking in small steps, requiring few proofs. To serve the needs of different readerships, a more accessible treatment of models as approximations is in Berk et al. (2019).

After working through further terminological difficulties in Rod Little's Section 3, we face in his Section 4 some language that is recognizable and agreeable to us: "...the slope of Y on X fitted to the entire population by least squares. This quantity exists regardless of whether the regression of Y on X is really linear, although its utility for summarization is weakened if the regression is highly nonlinear." We could have written this, so on this point there is no difference between his views and ours. It seems, therefore, that greater clarity and mutual understanding could be achieved, but, given human nature, this seems more likely in personal encounters scribbling on napkins than in isolated writing exercises.

Dalia Ghanem and Todd Kuffner make two interesting points, one about invariance to objective functions and another about causality.

The first point concerns objective functions called proper scoring rules, mentioned in Section 2.1 and described in Appendix A.2 of Part II, a common example being expected negative log-likelihoods.[6] Proper scoring rules are minimized when a model distribution agrees with the actual distribution. This implies for correctly specified models that the risks derived from proper scoring rules are all minimized by the correct model distribution. Things get interesting when Ghanem and Kuffner, citing Elliott, Ghanem and Krüger (2016) note that if the model $q(\vec{\boldsymbol{x}};\boldsymbol{\theta})$ of a 0-1 response $Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}$ is misspecified, then *different proper scoring rules lead to different best approximations* to the true conditional probability function $\mu(\vec{\boldsymbol{x}})=\boldsymbol{P}[Y=1|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}]$. At first glance, this would appear to be a qualitatively different condition for correct specification that just happens to parallel our notion of well-specification. A closer look reveals, however, that there exists a connection: For 0–1 responses, different choices of proper scoring rules amount to different choices of reweighting schemes, but with a twist, as will be shown next.

To start, we need to describe some basics of proper scoring rules for a 0–1 variable Y: A scoring rule is an objective function of the form $\mathcal{L}(Y,q)=YL_1(1-q)+(1-Y)L_0(q)$ for $q\in[0,1]$, where $L_1(1-q)$ and $L_0(q)$ are monotonic losses incurred by a probability forecaster who observes, respectively, $Y=1$ or $Y=0$ and guesses the value q for the unknown true $\mu=\boldsymbol{P}[Y=1]$. A scoring rule is "proper" if it is Fisher-consistent:

$$\begin{aligned}&\operatorname*{argmin}_q \boldsymbol{E}\big[\mathcal{L}(Y,q)\big]\\ &\quad=\operatorname*{argmin}_q\big[\mu L_1(1-q)+(1-\mu)L_0(q)\big]\\ &\quad=\mu.\end{aligned}\tag{2}$$

For smooth functions $L_{1/0}(\cdot)$, this condition implies that the derivative wrt q vanishes at $q=\mu$:

$$-\mu L_1'(1-\mu)+(1-\mu)L_0'(\mu)=0.$$

Hence $w(\mu)\triangleq L_1'(1-\mu)/(1-\mu)=L_0'(\mu)/\mu$ defines a weight function that characterizes the proper scoring rule (replacing μ with q in notation):

$$L_1'(1-q)=(1-q)w(q),\quad L_0'(q)=qw(q).\tag{3}$$

Special cases are the Bernoulli negative log-likelihood arising from $w(q)\propto 1/(q(1-q))$, and squared error or OLS, $(Y-q)^2$, arising from $w(q)\propto 1$.[7]

Turning to regression, let $q(\vec{\boldsymbol{X}};\boldsymbol{\theta})$ be a model (not assumed correct) for the true $\mu(\vec{\boldsymbol{X}})=\boldsymbol{P}[Y=1|\vec{\boldsymbol{X}}]$. Standard example: the linear logistic model, $q(\vec{\boldsymbol{X}};\boldsymbol{\theta})=\varphi(\boldsymbol{\theta}'\vec{\boldsymbol{X}})$, $\varphi(t)=1/(1+e^{-t})$. A proper scoring (PS)

[5] Szpiro et al. (2010) come close in their Section 4.2 where they discuss the fixed-$\boldsymbol{X}$ case. Their $\boldsymbol{\beta}_{\text{fixed}}$ denotes the same as our $\boldsymbol{\beta}(\boldsymbol{X})$.

[6] For more background, see Gneiting and Raftery (2007) and Buja, Stuetzle and Shen (2005).

[7] Squared loss turns into a proper scoring rules due to $Y^2=Y$: $(Y-q)^2=Y(1-q)+(1-Y)q$.

functional (Part II, Section 2.1) for $q(\vec{X};\boldsymbol{\theta})$ is obtained by

$$\begin{aligned}\boldsymbol{\theta}(\boldsymbol{P}) &= \operatorname*{argmin}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{P}}\big[\mathcal{L}\big(Y, q(\vec{X};\boldsymbol{\theta})\big)\big] \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{P}}\big[Y L_1\big(1-q(\vec{X};\boldsymbol{\theta})\big) \\ &\qquad + (1-Y) L_0\big(q(\vec{X};\boldsymbol{\theta})\big)\big].\end{aligned}$$

Such PS functionals are well-specified iff the model is correctly specified: $\mu(\vec{X}) = q(\boldsymbol{X};\boldsymbol{\theta}_0)$ ($\boldsymbol{P}$-a.s.) for some $\boldsymbol{\theta}_0$, hence $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{\theta}_0$. Next, we rewrite the (negative) gradient $\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{X}) = -\partial_{\boldsymbol{\theta}} \mathcal{L}(Y, q(\vec{X};\boldsymbol{\theta}))$ using the weight function $w(\cdot)$ from (3):

$$\begin{aligned}&\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{X}) \\ &\quad = \big(Y L_1'\big(1-q(\vec{X};\boldsymbol{\theta})\big) \\ &\qquad - (1-Y) L_0'\big(q(\vec{X};\boldsymbol{\theta})\big)\big)\partial_{\boldsymbol{\theta}} q(\vec{X};\boldsymbol{\theta}) \\ &\quad = \big(Y\big(1-q(\vec{X};\boldsymbol{\theta})\big) \\ &\qquad - (1-Y) q(\vec{X};\boldsymbol{\theta})\big) w\big(q(\vec{X};\boldsymbol{\theta})\big)\partial_{\boldsymbol{\theta}} q(\vec{X};\boldsymbol{\theta}) \\ &\quad = \big(Y - q(\vec{X};\boldsymbol{\theta})\big) w\big(q(\vec{X};\boldsymbol{\theta})\big)\partial_{\boldsymbol{\theta}} q(\vec{X};\boldsymbol{\theta}).\end{aligned} \tag{4}$$

Interpretation: Given a model $q(\vec{X};\boldsymbol{\theta})$ as approximant to $\mu(\vec{X})$, the best approximation in terms of a proper scoring rule $\mathcal{L}(Y, q)$ is found by solving the stationarity condition or estimating equation (EE) $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{X})] = \mathbf{0}$. In this EE, the proper scoring rule is represented by its weight function $w(q)$ defined in (3), which reweights according to the model value $q = q(\vec{X};\boldsymbol{\theta})$. In combination, one obtains a *regressor-dependent as well as parameter-dependent weight function* $w(q(\vec{X};\boldsymbol{\theta}))$. Thus different proper scoring rules differ in their reweighting inside the estimating equations, but reweighting is both $\vec{X}$- and $\boldsymbol{\theta}$-dependent.

It is now apparent that invariance to proper scoring rules is a form of invariance to reweighting, thereby sharing similarities with the reweighting diagnostic and the notion of well-specification proposed in Part II. We may next consider generalizing invariance to $\vec{X}$- and $\boldsymbol{\theta}$-dependent reweighting to characterize well-specification of EE functionals in general (Part II, Section 2.2). In doing so, we may permit the "weight" functions to have arbitrary signs and not be normalized:

- Let $\boldsymbol{\theta}(\boldsymbol{P})$ be defined by an estimating equation $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{X})] = \mathbf{0}$.
- Let $w(\vec{X};\boldsymbol{\theta}) \neq 0$ be a possibly parameter-dependent "weight" function.
- Define a reweighted score by

$$\tilde{\boldsymbol{\psi}}(\boldsymbol{\theta}; Y, \vec{X}) \stackrel{\Delta}{=} w(\vec{X};\boldsymbol{\theta})\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{X}).$$

- Define a reweighted EE functional $\tilde{\boldsymbol{\theta}}(\boldsymbol{P})$ by

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

The following is stated as usual without regularity conditions.

PROPOSITION. *The EE functionals $\boldsymbol{\theta}(\boldsymbol{P})$ and $\tilde{\boldsymbol{\theta}}(\boldsymbol{P})$ are well-specified for the same distributions $\boldsymbol{P}$. On these, the functionals have identical values*: $\boldsymbol{\theta}(\boldsymbol{P}) = \tilde{\boldsymbol{\theta}}(\boldsymbol{P})$.

PROOF. The EE functional $\boldsymbol{\theta}(\boldsymbol{P})$ is well-specified iff $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}_0; Y, \vec{X})|\vec{X}] = \mathbf{0}$ ($\boldsymbol{P}$-a.s.) for $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{P})$ (Proposition 3.3.3 of Part II). Because

$$\begin{aligned}&\boldsymbol{E}_{\boldsymbol{P}}\big[w(\vec{X};\boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{\theta}_0; Y, \vec{X})|\vec{X}\big] \\ &\quad = w(\vec{X};\boldsymbol{\theta}_0)\boldsymbol{E}_{\boldsymbol{P}}\big[\boldsymbol{\psi}(\boldsymbol{\theta}_0; Y, \vec{X})|\vec{X}\big]\end{aligned}$$

and $w(\vec{X};\boldsymbol{\theta}_0) \neq 0$, $\boldsymbol{\theta}(\boldsymbol{P})$ and $\tilde{\boldsymbol{\theta}}(\boldsymbol{P})$ are well-specified for the same distributions $\boldsymbol{P}$. The solutions of the two estimating equations are identical; hence $\boldsymbol{\theta}(\boldsymbol{P}) = \tilde{\boldsymbol{\theta}}(\boldsymbol{P}) = \boldsymbol{\theta}_0$. □

A practical consequence of the proposition is that when $\boldsymbol{\theta}(\boldsymbol{P})$ is misspecified for $\boldsymbol{P}$, then one may have $\boldsymbol{\theta}(\boldsymbol{P}) \neq \tilde{\boldsymbol{\theta}}(\boldsymbol{P})$, suggesting that nonconstancy under $\vec{X}$- and $\boldsymbol{\theta}$-dependent reweighting can be used as a diagnostic for misspecification. Consequently, nonconstancy across proper scoring rules can also be used for this purpose.[8] We are grateful to Ghanem and Kuffner for pointing us to proper scoring rules and giving us an opportunity to have a closer look at reweighted EE functionals as well.

Ghanem and Kuffner's second point concerns the connection of well-specification and causality (Part II, Section 3.4), and the notion of external validity of causal inference. According to Athey and Imbens (2017, Section 2.3), "...most concerns with external validity are related to treatment effect heterogeneity." It should generally be assumed that treatment effects vary (are heterogeneous) across different environments, even if treatment is randomized. To describe the situation, Ghanem and Kuffner use the multiple-environment notation of Peters et al. (2016) and consider a structural equation model (SEM) that includes

[8] Another use of proper scoring rules is proposed by Buja et al. (2005) for tailoring models to classification with asymmetric misclassification costs.

pretreatment covariates to capture the heterogeneity among environments. The differences between environments are then reduced to differences between covariate distributions to which the SEM is immune because it is assumed to capture the causal mechanisms correctly across environments. Quantities derived from the SEM, including the average treatment effect conditional on the covariates, will be well-specified, hence invariant across environments. In this sense, well-specification relates indeed to external validity.

Ghanem and Kuffner's discussion leads us to a follow-up thought: Questions of treatment effect heterogeneity arise also within a single study and are the subject of a growing literature, both for testing the presence of heterogeneity and for estimating its form as a function of pre-treatment covariates. See Athey and Imbens (2017, Section 10) for a selection of approaches, including nonparametric and tree-based function estimation, and Berk et al. (2020) for inference thereafter. To these approaches, we add the reweighting diagnostic of Part II as follows: Average treatment effects, whether marginal or conditional on a covariate location, form regression functionals, and as such they can be probed for well-specification using covariate-based reweighting. Misspecification amounts to heterogeneity of treatment effects. A fine point about the marginal average treatment effect is that it involves only a single regressor, the treatment variable. Yet it is meaningful to apply the diagnostic with weights from covariates other than the treatment variable. The reweighting diagnostic does not need to be limited to covariates used by the regression functional.

Alessandro Rinaldo, Ryan Tibshirani and Larry Wasserman (RTW for short) start by addressing common misinterpretations of regression results, even granting the correctness of the model. Indeed, many popular catch phrases used for interpretation imply causal dependence, which is misleading for most observational data. The teaching of regression should therefore emphasize that regression coefficients pertain to comparisons between *hypothetical* cases, such as pairs of humans whose heights (X_j) differ by 1 inch but whose weights (X_k) are the same. Such formulations are rhetorical losers compared to action metaphors such as "increase height by 1 inch" and "hold weight fixed," though these reveal themselves occasionally by their absurdity.

Yet, not all regressions are limited to correlation; some do call for careful causal thinking. Also, important questions will be ever more about causes, in science, public policy and business. The teaching of regression should, therefore, integrate critical thinking about the nature of causation and not just wave off the issue with the usual disclaimer that correlation is not causation. An example where reasoning about causality is difficult to avoid is provided by the LA Homeless data of Part I and II, where the regressor `PercVacant` is a potential candidate for intervention by public policies. The data alone do not allow us to infer that the slopes of `PercVacant` describe causal effects, but these data might nevertheless contain the most suggestive data-driven information available to guide intervention with public policies.

Moving on to misspecification and agreeing that this is the "more realistic" assumption, RTW give a gentle version of David Freedman's objection, wondering whether the parameters of a best approximation are meaningful for practitioners. This question is answered affirmatively in Part I, Section 10, with a universally valid interpretation of slopes for simple linear regression based on casewise and pairwise slopes. This interpretation can be leveraged for an intuitive way of teaching simple linear regression: consider line segments between all pairs of data points, obtain their slopes and form a weighted average. All that needs explaining is that pairs of points are more informative if they are distant in X, hence the weights. Then extend this interpretation to multiple regression by drawing on the language of adjustment: a multiple regression coefficient is the weighted average of pairwise slopes in a simple regression of the response on the regressor *linearly adjusted for all other regressors*. This avoids the phrase "at fixed levels of all other regressors" which assumes the true response surface to be linear.

RTW in their Section 3 examine assumption-lean regression in the sense of nonparametric regression, whereas our interest was in parametric regression without parametric assumptions. We can, however, point out that our reweighting diagnostic in Part II also represents a form of partial nonparametrics: We apply parametric models (actually: quantities of interest) to data localized by reweighting kernels, akin to local linear smoothers, but we do not use the localized models for estimating fitted values of surfaces. Rather, we use them to visualize a quantity of interest (often a slope) as a function of univariate kernels that localize over regressor dimensions of interest.

RTW proceed to discuss the question of regressor importance, generalizing what is often addressed by t-tests in linear models. An issue with regressor importance is that this notion has no absolute meaning. It only speaks to the predictive power of the regressor X_j above and beyond the other regressors $\vec{X}_{-j}$ in

the model.[9] Therefore, the null situation of a regressor X_j that is wholly *un*important for Y in the presence of $\vec{\boldsymbol{X}}_{-j}$ is conditional independence: $Y \perp\!\!\!\perp X_j | \vec{\boldsymbol{X}}_{-j}$. As RTW point out with reference to Shah and Peters (2018), this condition is difficult to test, unless $\vec{\boldsymbol{X}}_{-j}$ takes on finitely many values only. To get a heuristic sense for the difficulties, note that conditional independence implies $\mathrm{Cov}(g(Y), f(X_j)|\vec{\boldsymbol{X}}_{-j}) = 0$ (a.s.) for all measurable functions $f(X_j)$ and $g(Y)$ with second moments.[10] Equivalent is the following condition:

$$\begin{aligned} &\boldsymbol{E}\big[\big(g(Y) - \boldsymbol{E}\big[g(Y)|\vec{\boldsymbol{X}}_{-j}\big]\big) \\ &\quad \times \big(f(X_j) - \boldsymbol{E}\big[f(X_j)|\vec{\boldsymbol{X}}_{-j}\big]\big)h(\vec{\boldsymbol{X}}_{-j})\big] = 0, \end{aligned} \tag{5}$$

which additionally needs to hold for $h(\vec{\boldsymbol{X}}_{-j})$ in a set of functions whose linear span is dense in $L_2(\vec{\boldsymbol{X}}_{-j})$. Shah and Peters (2018) essentially propose a limited test for univariate variables based on the choices $f(X_j) = X_j$, $g(Y) = Y$ and $h(\vec{\boldsymbol{X}}_{-j}) = 1$, which illustrates how such tests may pick and choose aspects of the full null hypothesis of conditional independence. Of further interest is that (5) shows adjustment at work: $f(X_j) - \boldsymbol{E}[f(X_j)|\vec{\boldsymbol{X}}_{-j}]$ is $f(X_j)$ nonparametrically adjusted for $\vec{\boldsymbol{X}}_{-j}$. Conditions equivalent to (5) are obtained by adjusting $f(X_j)$ only, or $g(Y)$ only, or both (as shown). For practical adjustment on data, one may estimate/approximate conditional expectations $\boldsymbol{E}[\cdot|\cdot]$ with regression algorithms, $\mathcal{A}[\cdot|\cdot]$ in RTW's notation. For linear adjustment according to Part I, Sections 9ff, specialize the algorithm $\mathcal{A}$ to linear OLS and apply it to $f(X_j) = X_j$.

As for measures of regressor importance, (5) suggests a role for response transformations, as in ACE regression (Breiman and Friedman, 1985). There is no reason to expect that X_j is most strongly associated with the raw scale Y as opposed to some transformed scale $g(Y)$.

RTW mention several more ideas we are unable to address in this space but may inspire future work, most intriguingly those relating to conformal inference.

Nikki Freeman, Xiaotong Jiang, Owen Leete, Daniel Luckett, Teeranan Pokaprakarn and Michael Kosorok appreciate our analysis of the interplay of regressor randomness and misspecification but find themselves unable to fully resolve the questions raised by our framework. They bring to our attention that we may have focused too much on the inflammatory part of G.E.P. Box' famous quote (wrong models) and too little on the constructive part (useful models). Indeed, it is useful to think about what makes a model useful, and while some answers may be implicit in Part I and II, others should be added. A few indicators that in combination may render a model "useful" in the case of observational data are as follows:

- The model contains quantities of interest that remain interpretable under degrees of misspecification.
- For these quantities there exists assumption-lean inference.
- Some of the inferential conclusions are of interest, often some meaningful parameter estimates with strong p-values.

These answers flow from Part I and II. We do think that, for example, the analysis of the LA Homeless data in Part II, extended by the reweighting diagnostic, is interesting and possibly useful, though causal inference based on the data is doubtful. For contexts that require strict causal inference, the model must be chosen to allow proper estimation of causal effects. This last point can be violated in at least two ways:

- The model may be so viciously misspecified that its best approximation to the data may err in the direction of the estimated causal effect, an illustration of which is given in the discussion by Whitney et al. (see below).
- The model may include impermissible regressors such as other outcomes or exclude necessary regressors such as important confounders. Rules for selecting proper regressors are the subject of the DAG theory of causality (Pearl, 2009).

Obviously, the question of "useful models" is a much larger one, and we have given here only some tentative thoughts.

We agree with Freeman et al. when, toward the end of their Section 2, they allude to causal contexts with different requirements, some needing to establish an effect's direction only, others also its magnitude. Note, however, that causal studies often simplify the problem by using binary interventions, thereby ignoring potential nonlinear effects by design rather than flawed data analysis. The possibility of misspecification then resides largely in the adjustment for confounders (see Whitney et al.) and the modeling of effect heterogeneity (see Ghanem and Kuffner).

[9] $\vec{\boldsymbol{X}}_{-j} = (1, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)'$

[10] This condition is equivalent to conditional independence in most cases of interest, as when the sigma algebras of the X_j and Y spaces have countable bases, which is the case for all Polish spaces.

Freeman et al. quote McCullagh and Nelder (1983) who express a somewhat benign view of misspecification and model building. This quote, however, dates back to a time of innocence, before the emergence of vastly increased powers in methodology and computation that is available today. The present capabilities of algorithmic model searches and informal model explorations challenge the frequentist imagination, described earlier. A first step toward gaining some clarity would be to write down the anticipated "data analysis pipeline" (Freeman et al.'s useful term) in a document before seeing the data. Such a protocol would have several benefits: (1) It would reveal where data analysts take unanticipated steps; (2) it would offer an opportunity to automate some of the informal steps; and (3) if a high degree of automation were achieved, whole pipelines could be examined in simulation experiments wrt their effects on statistical inference, as illustrated by Simmons et al. (2011).

At the end of their Section 2, Freeman et al. address the interpretation of slopes under misspecification and note that "the functional could be interpreted as the average effect over the observed regressor distribution, which may still be a useful measure in this context." This informal shorthand can serve well in a causal context where "average effect" is the appropriate term, else "average difference in Y per difference in X for pairs of observations" could be used, which would become technically precise when preceded with the term "weighted ..."; see the rejoinder to Rinaldo et al. above and Part I, Section 10. A related meaning of "average effect" or "average slope" is used by econometricians who consider "average derivatives," and hence functionals $\boldsymbol{E}[\mu'(X)]$ where $\mu(x) = \boldsymbol{E}[Y|X = x]$ (Stoker, 1986; Newey et al., 2004, and references therein).

Freeman et al., in their third section, rightfully draw connections between semiparametric theory and our assumption-lean theory of regression functionals. The former is also a theory of low-dimensional functionals but focuses on their efficient estimation in the presence of nonparametric (infinite dimensional) nuisance "parameters." We did intentionally not touch semiparametrics in Part I and II because of the space needed to describe the effects of misspecification in the accessible case of linear OLS (Part I) and to construct an assumption-lean theory of regression functionals that encompasses the parameters of approximating parametric models and estimating equations with low-dimensional parameters (Part II). Semiparametric theory and our theory of regression functionals are concerned with different issues and neither solves the issues of the other, although they can help inform each other. Assumption-lean semiparametric theory has existed at least since Newey (1994). Also, the ingredients of the framework of Part I for linear OLS can be interpreted semiparametrically: the vector of slopes $\boldsymbol{\beta}(\boldsymbol{P})$ is a parametric functional, while the nonlinearity $\eta(\vec{\boldsymbol{X}})$ and the conditional variance $\sigma^2(\vec{\boldsymbol{X}})$ are nonparametric nuisances. The Pythagorean decompositions of Part I and II can be interpreted as deriving from orthogonal tangent spaces of semiparametric theory.

Importantly, even a semiparametric model such as the Cox model can be misspecified in the sense of Part II in that, for example, the true hazard function might differ by more than a proportionality between the low and high levels of a treatment. Such misspecification funnels the randomness of the regressors into the statistical variability of both the estimated treatment effect (= the parametric component) and the estimated baseline hazard function (= the nonparametric component). Finally, such misspecification could be detected with the reweighting diagnostic of Part II.

We will go into some more detail in the discussion of Whitney et al. below who also touch on issues of semiparametric inference.

David Whitney, Ali Shojaie and Marco Carone (WSC for short) focus vigorously on semiparametric theory, an area we intentionally avoided for reasons mentioned in the rejoinder to Freeman et al. above. Just the same, some of WSC's comments are accessible in the context of Parts I and II.

The first of their examples in their Section 2.1 concerns a too simple approximation leading to faulty conclusions: linear adjustment for a nonlinear confounder produces the wrong direction of an exposure effect. This can be seen as "faulty analysis" because of the stipulated causal context. Thus, if all models are misspecified to a degree, while some are useful, others may indeed be misleading. Here is a summary of WSC's example at a high level: There is an exposure variable X of interest in the presence of a complex confounder W. The nonlinear effects of the confounder and the joint regressor distribution (X, W) are constructed in such a way that introducing W linearly in the model results in—ironically—valid inference about the wrong direction and magnitude of the effect of the exposure variable. Essential to make the example work is the nature of the joint regressor distribution (X, W) (a horse-shoe shape), which reminds us of Jerald Lawless' remark:

"...I find it difficult to envision truly comprehensive analysis without considering both response models and covariate distributions."

The problem with the linear fit can be detected with various diagnostic tools, including our reweighting diagnostic (Part II). If, however, the confounder W were multivariate, the curse of dimensionality would make detection and estimation of nonlinear confounding effects progressively more difficult for increasing dimension of W. In addition, data dredging is in play again: if a complex confounding effect were anticipated before touching the data, valid inference for the correct slope of the exposure effect might be possible, but if detection is unanticipated, subsequent inference is post hoc and requires frequentist disclaimers, as discussed earlier.

In their Section 3, the authors go further and analyze nonparametric adjustment by fitting the functional form $Y \sim \theta X + g(W)$, $g(W) \in L_2(W)$ being an infinite-dimensional nuisance parameter.[11] The "exposure effect" $\theta = \theta(\boldsymbol{P})$ is a functional that can be written equivalently as follows:

$$\begin{aligned}\theta &= \frac{\boldsymbol{E}[Y(X-\pi(W))]}{\boldsymbol{E}[X(X-\pi(W))]} \\ &= \frac{\boldsymbol{E}[(Y-g_1(W))(X-\pi(W))]}{\boldsymbol{E}[X(X-\pi(W))]} \\ &= \frac{\boldsymbol{E}[(Y-g_2(W))(X-\pi(W))]}{\boldsymbol{E}[X(X-\pi(W))]},\end{aligned} \tag{6}$$

- where $\pi(W) = \boldsymbol{E}[X|W]$ is the propensity of exposure if $X \in \{0, 1\}$ is binary,
- $g_1(W) = E[Y|X = 0, W]$ is a "hack" assuming the functional form is correctly specified and exploiting $\boldsymbol{E}[Y|X = 0, W] = (\theta X + g(W))|_{X=0} = g(W)$,
- and, finally, $g_2(W)$ is part of the correct solution of the mixed parametric/nonparametric OLS problem: $\min_{\theta\in\mathbb{R}, g(W)\in L_2(W)} \boldsymbol{E}[(Y - \theta X - g(W))^2]$.

The identities follow from the fact that conditional expectations are orthogonal projections (idempotent and self-adjoint in $L_2(\boldsymbol{P})$) and render adjustment of Y optional and arbitrary as long as X is properly adjusted by $\pi(W)$. The reason for writing θ in the three different ways of (6) is that each inspires one of the estimators considered by WSC: $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$. The authors indicate that the first two estimators can have asymptotic bias problems, whereas the third permits a general semiparametric justification of model-robust inference. Although this observation is correct at this level of generality, the differences between the estimates can be removed by choosing particular nonparametric methods for estimating $\pi(W)$ and $g(W)$. The simplest way of removing the differences is by using an OLS projection onto a function space ("series estimator," Newey, 1994, p. 1372) spanned by a finite set of basis functions (e.g., B-splines where the analysis allows the number of knots to grow with the sample size). The identities (6) then carry over to estimates $\hat{\pi}(W)$ and $\hat{g}(W)$.

A second approach for removing differences in bias between the three estimates is based on "twicing" (Newey, Hsieh and Robins, 2004). It applies the fitting mechanism twice and adds the difference back to the first fit. As such it can be applied quite universally to nonparametric function estimation, including Parzen kernel smoothing and reproducing kernel smoothing. (It has no effect on projection-based function estimates due to their idempotence.) Both projection-based and twicing-based estimators can be shown to have a "Small Bias Property" (Newey et al., 2004, Section 3)[12] that enables semiparametric estimates of quantities of interest such as $\theta(\boldsymbol{P})$ to be $\sqrt{n}$-consistent even when the nonparametric function estimator converges at the optimal nonparametric rate. In summary, some of the dangers posed by intuitive but semiparametrically biased and inefficient estimators can be overcome by estimating the nuisance with suitable nonparametric methods.

The authors' Sections 2.2 and 4 address what they call "data coarsening," in particular right censored survival data, approached with the usual proportional hazards model. Misspecification difficulties are widely acknowledged despite the nonparametric nature of the conditional hazard function in the model: The assumptions of shared ratios $h_{\vec{x}}(t_2)/h_{\vec{x}}(t_1)$ across regressor space ($\vec{x}$) and shared ratios $h_{\vec{x}_1}(t)/h_{\vec{x}_2}(t)$ across time (t) both point to potential misspecifications. The authors cite complications under misspecification such as dependence of the regression functional $\boldsymbol{\theta}(\cdot)$ not only on the regressor distribution—familiar in our context—but also on the conditional time-to-event distribution as well as the censoring distribution. This is driven home by the simulation results shown in their Figure 2.

In their Section 4, the authors make a point which is entirely in the spirit of Part II: If a functional derived

[11] The authors notation is θ_0 and $g_0(W)$. We drop the subscripts and write $\hat{\theta}$ and $\hat{g}(W)$ for estimates.

[12] A better term would be "product bias property" because it is due to a product of two biases in the third numerator of (6), which in combination can speed up convergence to a rate faster than $1/\sqrt{n}$.

from a model has opaque behavior under misspecification, revise the functional. In the context of a misspecified Cox model, they propose targeting a time average of the true log-hazard ratio, which has intuitive meaning as well as good estimation properties. This example is a nice illustration of one of the main messages of Part II: Focus on *quantities of interest rather than models*. Contrary to the authors' statements at the beginnings of their Sections 3 and 4, the framework of Part II does *not* require the quantities of interest to be model-based. To the opposite, Part II frees itself explicitly from this limitation and presents a theory of quantities of interest without reference to models, which is what the authors intended with their term "model agnostic." For ease of exposition only, we illustrate how this theory specializes to familiar quantities that are parameters of approximating models (ML and PS functionals).

We note additional complexity not discussed by WSC in the context of the Cox model for censored outcomes: The authors' proposed semiparametric estimator of "treatment effect" relies on an assumption of independent censoring within treatment arm, which allows them to avoid having to explicitly estimate the censoring mechanism. It would, however, be desirable to treat the censoring mechanism as a genuine nuisance because it is study-specific and not of scientific interest. This could be done by reweighting each person-time contribution by the corresponding inverse conditional probability of remaining uncensored given treatment arm. As a result, the limiting MPLE functional for the exposure effect would be freed of its dependence on the censoring mechanism under the authors' assumptions. These issues, however, become more complicated in scenarios that may be closer to what is often seen in practice:

- When the Cox working model involves several covariates rather than a single binary regressor, defining an alternative parameter target in a spirit similar to the authors' θ_{**} may be more challenging. Under misspecification, the hazard function $h_{\vec{x}}(t)$ may vary with the location $\vec{x}$ in a multivariate regressor space. The authors' time-average of $\beta(t) = \log(h_{x=1}(t)/h_{x=0}(t))$ for a single binary regressor $x \in \{0, 1\}$ is then no longer available. The weighted MPLE, on the other hand, may still provide a viable solution.
- When the censoring mechanism depends on numerous fully observed time-varying covariates, correcting for the censoring mechanism may require estimating a model for censoring given the covariate process. This applies to either approach, the authors' or the weighted analysis mentioned above. Thus issues of misspecification, now of the censoring model, equally affect both approaches for defining a meaningful target parameter under a misspecified model, either as the limit of the weighted MPLE or the alternative functional defined by the authors. Neither can truly escape the issue of misspecification.

Finally, misspecification of a censoring model is relevant mostly in so far as it affects the quantity of primary interest, often the regression slope of an exposure variable. Because of the intended interpretation as a causal effect, it is desirable that the quantity does not depend on the covariate distribution (Part II, Section 3.4). This, of course, is the definition of well-specification of a quantity of interest. It would therefore be useful to examine whether the reweighting diagnostic for well-specification of Part II could be applied in the context of a Cox model with censored outcomes.

Davison, Koch and Koh are first of all to be thanked for finding an error in our Table 4 of Part I, which allowed us to correct it before publication. We appreciate their skills as data detectives. In the substance, their discussion is the most searing criticism of our articles from the perspective of traditional model building: perform EDA before modeling and apply diagnostics afterward. If the latter indicate discrepancies, then the "likely impact on the conclusions needs to be assessed, and the benefit of dealing with it weighed against the cost of doing so." While leaving unspecified what such assessing and weighing might be, the quote indicates an awareness that model modification based on diagnostics has inferential costs. The same, however, is true for the initial EDA. If the model choice is influenced by an exploratory peek at the data, the inferential sin has already been committed. Once again we need to appeal to the frequentist imagination: What modeling decisions might we have made with a similar peek at alternative datasets drawn from the same population? To answer this question, EDA procedures and resulting data analytic decisions need to be formalized in protocols that can be simulated, if not analyzed theoretically.[13] Even in clinical trials, where statistical inference is at its most rigorous, agencies are pushing to tighten protocols (EMA, FDA, 2017).

[13] Simmons et al. (2011) give an example how this could play out. They formalized and simulated some of the lax practices prevalent in social sciences to illustrate their detrimental effects on statistical inference.

Davison and coauthors express astonishment at the implications of our articles because they seem to "run counter to at least a century of statistical practice." This practice — EDA before and diagnostics after model fitting—is what most of us have been taught and what we have been teaching in turn. The assumption of correct model specification, however, has *not* been uniformly adopted by statisticians, and concern with misspecification has been present in at least two traditions, those of econometrics (White, 1980a, 1980b) and of estimating equations (e.g., Godambe and Thompson, 1984; Boos, 1992). Indeed, Godambe and Thompson (1984) begin by explaining the two meanings of the term "parameter." One "characterizes some 'interesting' aspect of the distribution such as its mean," while the other "derives its meaning from a probabilistic model." Just as the mean is meaningful for distributions other than Gaussians, the linear trend produced by linear OLS is meaningful for joint distributions other than those of linear models. This is another way of describing the meaning of "misspecification," which in this sense certainly has been part of statistical practice.

Davison and coauthors seem to argue next that regression can only provide approximations for the observed points in regressor space.[14] This argument would clearly be unreasonable because it precludes the use of regression for prediction at not previously observed regressor locations. If the authors are concerned with the important and difficult problem of extrapolation, then there is good news. As indicated in the rejoinder to Jerald Lawless, the reweighting methodology of Part II can be applied to prediction functionals to detect serious extrapolation problems.

It is only natural that Davison and coauthors also argue in favor of fixed-$\boldsymbol{X}$ theory and the conditional target $\boldsymbol{\beta}(\boldsymbol{X})$ as opposed to the population target $\boldsymbol{\beta}(\boldsymbol{P})$, calling the latter "inestimable." But $\boldsymbol{\beta}(\boldsymbol{P})$ is consistently estimated by the linear OLS estimator $\boldsymbol{\beta}(\hat{\boldsymbol{P}}_N)$. Insisting on the fixed-$\boldsymbol{X}$ treatment conceals from the frequentist imagination the fact that in observational studies, hypothetical datasets $(\boldsymbol{y}, \boldsymbol{X})$ drawn from the same population not only differ in $\boldsymbol{y}$ but in $\boldsymbol{X}$ as well. The full variability in $\boldsymbol{\beta}(\hat{\boldsymbol{P}}_N) - \boldsymbol{\beta}(\boldsymbol{P})$ across datasets includes a contribution from $\boldsymbol{\beta}(\boldsymbol{X}) - \boldsymbol{\beta}(\boldsymbol{P})$ arising from the regressor randomness in the presence of misspecification. This has nothing to do with extrapolation and everything with a full accounting of the sources of sampling variability.

In their Section 2, the authors argue that "divergences between the assumed and true models that can easily be detected are not of interest," because then "the fit [should be] improved so that the only remaining divergences ... are on the borderline of detectability." This principle is a recipe for reproducibility disasters. Even if a "divergence" is detected, there is often a multitude of remedies, such as constructing more regressors from the existing ones and/or transforming the response and the regressors in some manner. Different data analysts might apply different remedies. Even worse, in the frequentist imagination the same data analyst might have used different remedies on different datasets drawn from the same population. As a consequence, gifted data analysts might well be significant contributors to unreproducible empirical results.

We also need to reiterate our earlier argument that there exist situations where we decide to use a model even if we are able to detect misspecification, one possible reason being the need to commit to a protocol in the interest of reproducibility. We therefore strongly argue against the authors' supposition that misspecification is of interest only when it is on the borderline of detectability.

Against the authors' Section 3, it needs to be restated that the ***RAV*** test is *not* a powerful general-purpose misspecification test. It does not test for nonlinearities or heteroskedasticities in general. Instead, it tests whether these misspecifications cause detectable discrepancies between the true standard error and the one derived from linear models theory. As illustrated in Part I, Section 11.6, there exist misspecifications for which the model-trusting standard error of linear models theory is just fine. A fortuitous illustration of this message is provided by the authors' simulation examples. The analytical specifications of their scenarios allow us to calculate the true values of ***RAV*** to high precision, and the results are as follows: In the scenario of nonlinearity alone with homoskedastic noise ($\gamma = 0.7,\ \delta = 0.0,\ N = 50$) we have $\boldsymbol{RAV} = 1.04$, and in the scenario of heteroskedasticity alone with absent nonlinearity ($\gamma = 0.0,\ \delta = 0.6,\ N = 50$) we have $\boldsymbol{RAV} = 1.09$.[15] Taking roots of these values, we see that asymptotically the true standard error deviates

[14] They use the expression "the region $\mathcal{X}$ in which values of $\vec{\boldsymbol{x}}$ are known," but this terminology is mathematically undefined: Is $\mathcal{X}$ the set of observed regressor locations $\{\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_N\}$, or its convex hull, or a surrounding Mahalanobis ellipsoid appropriate for linear regression?

[15] For $N = 100$ and 200, the misspecifications are of lesser magnitude and the ***RAV*** values even closer to 1.

from the linear models standard error by just 2% in the nonlinear scenario, and by just 4.5% in the heteroskedastic scenario. These discrepancies are so minor as to be practically negligible, hence the authors produced illustrations of situations where the standard errors of linear models theory are fine. Because these scenarios form approximate null hypotheses of the ***RAV*** test, we should expect rejection probabilities near the nominal level α. Deeply gratifying to us is therefore what the authors report: Performing ***RAV*** tests at $\alpha = 0.05$, their simulations produce estimates of rejection probabilities in the vicinity of 0.07. The authors should rejoice also as there is no reason in their scenarios to mistrust the traditional standard error estimates of the linear slope.

A note of caution is in order regarding the limited insights to be gained from single-regressor examples. Our figures in Part I are of course also drawn for single regressors, but their messages are easily extended to $p > 1$; furthermore, we give a warning in Section 4.2 about difficulties arising from misspecification when $p > 1$. The issue is less with detection of misspecification and more with the multitude of choices for detecting and fixing it (graphical displays, misspecification tests, inclusion of nonlinear and interaction terms, Box-Cox transformations, additive models, ACE regression, tree-based methods, random forests, boosting, kernelizing, ...). The wealth of modeling choices available today and its unfettered use should surely count as "researcher degrees of freedom" (Simmons et al., 2011). Those who are rightfully weary of free-wheeling model building and desire valid inference under misspecification should obtain model-trusting as well as assumption-lean standard errors and compare them. There is a difference between using model-trusting standard errors with justification and using them blindly.

In their Section 4, the authors give a useful tutorial on the connection between bootstrap and sandwich estimators, with differences in notation: Where we write $\boldsymbol{\theta}(\boldsymbol{P})$, they write $t(F)$. Their tutorial, however, is about $N \to \infty$, whereas our connection between pairs bootstrap and sandwich estimators addresses the resample size $M \to \infty$ for *fixed sample size* N. See Part I, Section 8.2 and Part II, Section 8.1: *The plug-in/sandwich estimator is the limit of the M-of-N bootstrap estimator as* $M \to \infty$ *for fixed* N. Compared to the authors' tutorial, this fact is so simple that its precise formulation is its own proof. It might also be new as the authors' tutorial does not address this case.

In their Section 5, the authors discuss designed experiments, not the type of data addressed by our articles, but interesting nevertheless. We agree that here the pairs bootstrap does not offer a natural approach to inference, yet the issue of misspecification and the need for assumption-lean inference persists. Even if the design points $\vec{\boldsymbol{x}}_i$ chosen by experimenters do not represent a population, the theory of designed experiments operates formally with a representation of the design as an empirical measure $\hat{\boldsymbol{P}}_{\vec{\boldsymbol{X}}} = \frac{1}{N}\sum_i \delta_{\vec{\boldsymbol{x}}_i}$. Reweighting could still be used to detect misspecification of regression functionals in the sense of Part II. As for assumption-lean inference, the pairs bootstrap can indeed have problems due to a nonzero probability of generating singular design matrices. A partial solution could be to either use sandwich estimators or M-of-N bootstrap estimators with very large resample size M, which reduces the probability of singular designs. A clean computational solution has been proposed by Koller and Stahel (2017) who devise highly efficient ways of nonsingular sampling. An alternative solution is the multiplier bootstrap, which is also assumption-lean and does not suffer the problem of generating singular designs.

We conclude by thanking the discussants once again for their thoughtful and sometimes critical arguments. It has been a great pleasure for us to engage in a spirited debate. We are especially grateful to the Editor, Cun-Hui Zhang, for organizing this discussion.

ACKNOWLEDGMENTS

A. Buja and L. Zhao were supported in part by NSF Grant DMS-10-07657 and DMS-1310795. E. George was supported in part by NSF Grant DMS-14-06563.

REFERENCES

ADAM, D. (2019). Psychology's reproducibility solution fails first test. *Science* **364** 813. 10.1126/science.364.6443.813.

ARONOV, P. M. and MILLER, B. T. (2019). *Foundations of Agnostic Statistics*. Cambridge Univ. Press, Cambridge.

ATHEY, S. and IMBENS, G. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments* 1 73–140. Elsevier, Amsterdam.

AZRIEL, D., BROWN, L. D., SKLAR, M., BERK, R., BUJA, A. and ZHAO, L. (2016). Semi-supervised linear regression. Available at arXiv:1612.02391.

BERK, R., BUJA, A., BROWN, L., GEORGE, E., KUCHIBHOTLA, A. K., SU, W. and ZHAO, L. (2019). Assumption lean regression. *Amer. Statist.* 10.1080/00031305.2019.1592781.

BERK, R., OLSON, M., BUJA, A. and OUSS, A. (2020). Using recursive partitioning to find and estimate heterogeneous treatment effects in randomized clinical trials. *J. Exp. Criminol.*. To appear. Available at arXiv.org/abs/1807.04164.

BOOS, D. D. (1992). On generalized score tests. *Amer. Statist.* **46** 327–333.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. MR0803258

BUJA, A., STUETZLE, W. and YI, S. (2005). Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications. Unpublished manuscript. Available at www-stat.wharton.upenn.edu/~buja.

CANTONI, E. and RONCHETTI, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96** 1022–1030. MR1947250

DAVIES, L. (2014). *Data Analysis and Approximate Models*: *Model Choice*, *Location-Scale*, *Analysis of Variance*, *Nonparametric Regression and Image Analysis*. *Monographs on Statistics and Applied Probability* **133**. CRC Press, Boca Raton, FL. MR3241514

ELLIOTT, G., GHANEM , D. and KRÜGER, F. (2016). Forecasting conditional probabilities of binary outcomes under misspecification. *The Review of Economics and Statistics* **98** 742–755.

EMA, FDA (2017). ICH E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials. www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials, www.regulations.gov/docket?D=FDA-2017-D-6113.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GODAMBE, V. P. and THOMPSON, M. E. (1984). Robust estimation through estimating equations. *Biometrika* **71** 115–125. MR0738332

HARTMAN, N. (2014). Who really found the Higgs Boson. https://getpocket.com/explore/item/who-really-found-the-higgs-boson.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *Chance* **18** 40–47. MR2216666

KOLLER, M. and STAHEL, W. A. (2017). Nonsingular subsampling for regression S estimators with categorical predictors. *Comput. Statist.* **32** 631–646. MR3656977

KUCHIBHOTLA, A. K., BROWN, L. D. and BUJA, A. (2018a). Model-free study of ordinary least squares linear regression. Available at arXiv:1809.10538.

KUCHIBHOTLA, A. K., BROWN, L. D., BUJA, A., GEORGE, E. I. and ZHAO, L. (2018b). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. Available at arXiv:1802.05801.

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342

MCCARTHY, D., ZHANG, K., BROWN, L. D., BERK, R., BUJA, A., GEORGE, E. I. and ZHAO, L. (2018). Calibrated percentile double bootstrap for robust linear regression inference. *Statist. Sinica* **28** 2565–2589. MR3839874

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized linear models*. Chapman and Hall, London.

NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237

NEWEY, W. K., HSIEH, F. and ROBINS, J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* **72** 947–962. MR2051442

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. MR3557186

PITKIN, E., BERK, R., BROWN, L., BUJA, A., GEORGE, E., ZHANG, K. and ZHAO, L. (2013). Improved precision in estimating average treatment effects. Available at arXiv:1311.0291.

SHAH, R. and PETERS, J. (2018). The hardness of conditional independence testing and the generalised covariance measure. Available at arXiv:1804.07203.

SIMMONS, J. P., NELSON, L. D. and SIMONSOHN, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22** 1359–1366.

SZPIRO, A. A., RICE, K. M. and LUMLEY, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *Ann. Appl. Stat.* **4** 2099–2113. MR2829948

STEINBERGER, L. and LEEB, H. (2018). Conditional predictive inference for high-dimensional stable algorithms. Available at arXiv:1809.01412v1.

STOKER, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481. MR0868152

WHITE, H. (1980a). Using least squares to approximate unknown regression functions. *Internat. Econom. Rev.* **21** 149–170. MR0572464

WHITE, H. (1980b). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027