Statistical Science
2019, Vol. 34, No. 4, 523–544
https://doi.org/10.1214/18-STS693

Models as Approximations I: Consequences Illustrated with Linear Regression[1]

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang and Linda Zhao

Abstract. In the early 1980s, Halbert White inaugurated a "model-robust" form of statistical inference based on the "sandwich estimator" of standard error. This estimator is known to be "heteroskedasticity-consistent," but it is less well known to be "nonlinearity-consistent" as well. Nonlinearity, however, raises fundamental issues because in its presence regressors are not ancillary, hence cannot be treated as fixed. The consequences are deep: (1) population slopes need to be reinterpreted as statistical functionals obtained from OLS fits to largely arbitrary joint x-y distributions; (2) the meaning of slope parameters needs to be rethought; (3) the regressor distribution affects the slope parameters; (4) randomness of the regressors becomes a source of sampling variability in slope estimates of order $1/\sqrt{N}$; (5) inference needs to be based on model-robust standard errors, including sandwich estimators or the x-y bootstrap. In theory, model-robust and model-trusting standard errors can deviate by arbitrary magnitudes either way. In practice, significant deviations between them can be detected with a diagnostic test.

Key words and phrases: Ancillarity of regressors, misspecification, econometrics, sandwich estimator, bootstrap.

1. INTRODUCTION

Halbert White's basic sandwich estimator of standard error for OLS can be described as follows: In a

Andreas Buja is the Liem Sioe Liong/First Pacific Company Professor of Statistics, Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104-6340, USA. Lawrence Brown was the Miers Busch Professor of Statistics, Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104-6340, USA. Richard Berk is Professor of Criminology and Statistics, Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104-6340, USA. Edward George is the Universal Furniture Professor of Statistics, Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104-6340, USA. Emil Pitkin is Lecturer and Research Scholar in Statistics, Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104-6340, USA.

Mikhail Traskin was Assistant Professor of Statistics, work done while with the Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104-6340, USA. Kai Zhang is Associate Professor, Department of Statistics & Operations Research, University of North Carolina at Chapel Hill, 306 Hanes Hall, CB#3260, Chapel Hill, North Carolina 27599-3260, USA. 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).

[1]Discussed in 10.1214/19-STS722, 10.1214/19-STS723, 10.1214/19-STS724, 10.1214/19-STS725, 10.1214/19-STS726, 10.1214/19-STS746, 10.1214/19-STS747, 10.1214/19-STS748, 10.1214/19-STS756; rejoinder at 10.1214/19-STS762.

linear model with regressor matrix $\mathbf{X}_{N\times(p+1)}$ and response vector $\boldsymbol{y}_{N\times 1}$, start with the familiar derivation of the covariance matrix of the OLS coefficient estimate $\hat{\boldsymbol{\beta}}$, but allow heteroskedasticity, $\boldsymbol{V}[\boldsymbol{y}|\mathbf{X}] = \boldsymbol{D}$ diagonal:

$$\begin{aligned}\boldsymbol{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \boldsymbol{V}\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}|\mathbf{X}\big] \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{D}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned} \tag{1}$$

The last expression has the characteristic "sandwich" form, $(\mathbf{X}'\mathbf{X})^{-1}$ forming the "bread" and $\mathbf{X}'\boldsymbol{D}\mathbf{X}$ the "meat." Although this sandwich formula does not look actionable for standard error estimation because the variances $\boldsymbol{D}_{ii} = \sigma_i^2$ are not known, White showed that (1) can be estimated asymptotically correctly. If one estimates σ_i^2 by squared residuals r_i^2, each r_i^2 is not a good estimate, but the averaging implicit in the "meat" provides an asymptotically valid estimate:[1]

$$\hat{\boldsymbol{V}}_{\text{sand}}[\hat{\boldsymbol{\beta}}] \triangleq (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\boldsymbol{D}}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}, \tag{2}$$

where $\hat{\boldsymbol{D}}$ is diagonal with $\hat{\boldsymbol{D}}_{ii} = r_i^2$. Standard error estimates are obtained by $\hat{\boldsymbol{SE}}_{\text{sand}}[\hat{\beta}_j] = \hat{\boldsymbol{V}}_{\text{sand}}[\hat{\boldsymbol{\beta}}]_{jj}^{1/2}$. They are asymptotically valid even if the responses are heteroskedastic, hence the term "Heteroskedasticity-Consistent Covariance Matrix Estimator" in the title of one of White's (1980b) famous articles.

Lesser known is the following deeper result in one of White's (1980a, p. 162–163) less widely read articles: the sandwich estimator of standard error is asymptotically correct even in the presence of nonlinearity:[2]

$$\boldsymbol{E}[\boldsymbol{y}|\mathbf{X}] \neq \mathbf{X}\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta}. \tag{3}$$

The term "heteroskedasticity-consistent" is an unfortunate choice as it obscures the fact that the same estimator of standard error is also "nonlinearity-consistent" *when the regressors are treated as random*. The sandwich estimator of standard error is therefore "model-robust" not only against second-order model violations but first-order violations as well. Because of the relative obscurity of this important fact we will pay considerable attention to its implications. In particular, we will show how *nonlinearity* "conspires" with *randomness of the regressors*:

(1) to make slopes dependent on the regressor distribution and

(2) to generate sampling variability, even in the absence of noise in the response.

For an intuitive grasp of these effects, the reader may peruse Figure 2 for effect (1) and Figure 4 for effect (2).[3]

From the sandwich estimator (2), the usual *model-trusting* estimator is obtained by collapsing the sandwich form using homoskedasticity, $\hat{\boldsymbol{D}} = \hat{\sigma}^2\boldsymbol{I}$:

$$\begin{aligned}\hat{\boldsymbol{V}}_{\text{lin}}[\hat{\boldsymbol{\beta}}] &\triangleq (\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2, \\ \hat{\sigma}^2 &= \|\boldsymbol{r}\|^2/(N-p-1).\end{aligned}$$

This yields finite-sample unbiased squared standard error estimators $\hat{\boldsymbol{SE}}^2_{\text{lin}}[\hat{\beta}_j] = \hat{\boldsymbol{V}}_{\text{lin}}[\hat{\boldsymbol{\beta}}]_{jj}$ if the model is first- and second-order correct: $\boldsymbol{E}[\boldsymbol{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$ (linearity) and $\boldsymbol{V}[\boldsymbol{y}|\mathbf{X}] = \sigma^2\boldsymbol{I}_N$ (homoskedasticity). Assuming distributional correctness (Gaussian errors), one obtains finite-sample correct tests and confidence intervals.

The corresponding tests and confidence intervals based on the sandwich estimator have only an asymptotic justification, but their asymptotic validity holds under much weaker assumptions. In fact, it may rely on no more than the assumption that the rows $(y_i, \vec{\boldsymbol{x}}_i')$ of the data matrix $(\boldsymbol{y}, \mathbf{X})$ are i.i.d. samples from a joint multivariate distribution subject to some technical conditions. Thus sandwich-based theory provides asymptotically correct inference that is *model-robust*. The question then arises what model-robust inference is about: When no model is assumed, *what are the parameters*, and *what is their meaning*?

Discussing these questions is a first goal of this article. An established answer is that parameters can be reinterpreted as *statistical functionals* $\boldsymbol{\beta}(\boldsymbol{P})$ defined on a large nonparametric class of joint distributions $\boldsymbol{P} =$

[1] This sandwich estimator is only the simplest version of its kind. Other versions were examined, for example, by MacKinnon and White (1985) and Long and Ervin (2000). Some forms are pervasive in Generalized Estimating Equations (GEE; Liang and Zeger, 1986, Diggle et al., 2002) and in the Generalized Method of Moments (GMM; Hansen, 1982, Hall, 2005).

[2] The term "nonlinearity" is meant in the sense of first order model misspecification. A different meaning of "nonlinearity," *not* intended here, occurs when the regressor matrix $\mathbf{X}$ contains multiple columns that are functions (products, polynomials, B-splines, ...) of underlying independent variables. One needs to distinguish between "regressors" and "independent variables": Multiple regressors may be functions of one or more independent variable(s).

[3] A more striking illustration of effect (2) in the form of an animation is available to users of the **R** *Language* (2008) by executing the following line of code:

```
source("http://stat.wharton.upenn.edu/~buja/
  src-conspiracy-animation2.R")
```

$\boldsymbol{P}(dy, d\vec{\boldsymbol{x}})$ through best approximation (Section 3), sometimes called "projection." The sandwich estimator produces then asymptotically correct standard errors for the slope functionals $\beta_j(\boldsymbol{P})$ (Section 5). Vexing is the question of the meaning of slopes in the presence of nonlinearity as the standard interpretations no longer apply. We will propose interpretations that draw on the notions of *casewise* and *pairwise slopes* after linear adjustment (Section 10).

A second goal of this article is to discuss why the regressors should be treated as random. Based on an ancillarity argument, model-trusting theories tend to condition on the regressors and hence treat them as fixed (Cox and Hinkley, 1974, p. 32f; Lehmann and Romano, 2008, p. 395ff). However, it will be shown that under misspecification ancillarity of the regressors is violated (Section 4). Here are some implications:

- Population parameters $\boldsymbol{\beta}(\boldsymbol{P})$, now interpreted as statistical functionals, depend on the distribution of the regressors. Thus it matters where the regressors fall. The reason is intuitive: When models are approximations, it matters where the approximation is made; see Figure 2.
- A natural intuition fails, caused by misleading terminology: Nonlinearity — sometimes called "model bias"—does *not* primarily cause bias in estimates $\boldsymbol{\beta}(\hat{\boldsymbol{P}})$. It causes sampling variability of order $N^{-1/2}$, thereby rivaling error/noise as a source of sampling variability (Section 6).
- A second intuition fails: While it is correct that an inference guarantee conditional on the regressors implies a marginal inference guarantee, this principle is inapplicable because the premise is false—*under misspecification, there is no inference guarantee conditional on the regressors.* The reason is that inference theories that treat regressors as fixed are incapable of correctly accounting for misspecification.

All three implications hold in great generality, but in this article they will be worked out for OLS linear regression to achieve the greatest degree of lucidity.

A third goal of this article is to argue in favor of the "x–y bootstrap" which resamples observations $(\vec{\boldsymbol{x}}_i', y_i)$. The better known "residual bootstrap" resamples residuals r_i and thereby assumes a linear response surface and exchangeable errors. There exists theory to justify both (e.g., Freedman, 1981, and Mammen, 1993), but only the x–y bootstrap is model-robust and solves the same problem as the sandwich estimator.[4] In Part II of this two-part series of articles, it will be shown that the sandwich estimator is a limiting case of the x–y bootstrap.

A fourth goal of this article is to practically (Section 2) and theoretically (Section 11) compare model-robust and model-trusting estimators of standard error in the case of OLS linear regression. To this end, we define a ratio of asymptotic variances—"***RAV***" for short—that describes the discrepancies between the two standard errors in the asymptotic limit.

A fifth goal is to estimate the ***RAV*** for use as a test statistic. We derive an asymptotic null distribution to test for model deviations that invalidate the usual standard error of a specific coefficient. The resulting "misspecification test" differs from other such tests in that it answers the question of discrepancies among standard errors directly and separately for each coefficient (Section 12).

A final goal is to briefly discuss issues with sandwich estimators (Section 13): They can be inefficient when models are correctly specified. We additionally point out that they are nonrobust to heavy tails in the joint x–y distribution. To make sense of this observation, the following distinctions are needed: (1) classical robustness to heavy tails is distinct from model robustness to first- and second- order model misspecifications; (2) at issue is not robustness (in either sense) of parameter estimates but of standard errors. It is the latter we examine here.

Throughout we use precise notation for clarity, yet this article is not very technical. Many results are elementary, not new, and stated without regularity conditions. Readers may browse the tables and figures and read associated sections that seem most germane. Important notations are shown in boxes.

The present article is limited to OLS linear regression, both for populations and for data. The case permits explicit calculations and lucid interpretations. Part II analyzes the notion of regression in a general and model-free way.

The idea that models are approximations, and hence generally "misspecified" to a degree has a long history, most famously expressed by Box (1979). We prefer to quote Cox (1995): "it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization."

[4]Note David Freedman's (1981) surprise when he inadvertently discovered the same assumption-lean validity of the x–y bootstrap (ibid. top of p. 1220).

TABLE 1
LA Homeless data: *Comparison of standard errors*. *See also Table* 2 *in the Appendix* (Buja et al., 2019) *for the Boston Housing Data*

	$\hat{\beta}_j$	$\boldsymbol{SE}_{\text{lin}}$	$\boldsymbol{SE}_{\text{boot}}$	$\boldsymbol{SE}_{\text{sand}}$	$\frac{\boldsymbol{SE}_{\text{boot}}}{\boldsymbol{SE}_{\text{lin}}}$	$\frac{\boldsymbol{SE}_{\text{sand}}}{\boldsymbol{SE}_{\text{lin}}}$	$\frac{\boldsymbol{SE}_{\text{sand}}}{\boldsymbol{SE}_{\text{boot}}}$	t_{lin}	t_{boot}	t_{sand}
`Intercept`	0.760	22.767	16.505	16.209	**0.726**	**0.712**	0.981	0.033	0.046	0.047
`MedianIncome ($K)`	−0.183	0.187	0.114	0.108	**0.610**	**0.576**	0.944	−0.977	−1.601	−1.696
`PercVacant`	4.629	0.901	1.385	1.363	**1.531**	**1.513**	0.988	5.140	3.341	3.396
`PercMinority`	0.123	0.176	0.165	0.164	0.937	0.932	0.995	0.701	0.748	0.752
`PercResidential`	−0.050	0.171	0.112	0.111	**0.653**	**0.646**	0.988	−0.292	−0.446	−0.453
`PercCommercial`	0.737	0.273	0.390	0.397	**1.438**	**1.454**	1.011	2.700	1.892	1.857
`PercIndustrial`	0.905	0.321	0.577	0.592	**1.801**	**1.843**	1.023	2.818	1.570	1.529

The history of inference under misspecification can be traced to Cox (1962), Eicker (1963), Berk (1966, 1970), Huber (1967), before being systematically elaborated by White's articles (1980a, 1980b, 1981, 1982, among others), capped by a monograph (White, 1994). A wide-ranging discussion by Wasserman (2011) calls for "Low Assumptions, High Dimensions." A book by Davies (2014) elaborates the idea of adequate models for a given sample size. We, the present authors, got involved with this topic through our work on post-selection inference (Berk et al., 2013) because the results of model selection should certainly not be assumed to be "correct." We compared the obviously model-robust standard errors of the x–y bootstrap with the usual ones of linear models theory and found the discrepancies illustrated in Section 2. Attempting to account for these discrepancies became the starting point of the present article.

2. DISCREPANCIES BETWEEN STANDARD ERRORS ILLUSTRATED

Table 1 shows regression results for a dataset consisting of a sample of 505 census tracts in Los Angeles that has been used to relate the local number of homeless (Y) to covariates for demographics and building usage (Berk et al., 2008).[5] We do not intend a careful modeling exercise but show the raw results of linear regression to illustrate the degree to which discrepancies can arise among three types of standard errors: $\boldsymbol{SE}_{\text{lin}}$ from linear models theory, $\boldsymbol{SE}_{\text{boot}}$ from the x–y bootstrap ($N_{\text{boot}} = 100{,}000$) and $\boldsymbol{SE}_{\text{sand}}$ from the sandwich estimator (according to MacKinnon and White's (1985) HC2 proposal). Ratios of standard errors that are far from +1 are shown in bold font.

The ratios $\boldsymbol{SE}_{\text{sand}}/\boldsymbol{SE}_{\text{boot}}$ show that the sandwich and bootstrap estimators are in good agreement. Not so for the linear models estimates: we have $\boldsymbol{SE}_{\text{boot}}$, $\boldsymbol{SE}_{\text{sand}} > \boldsymbol{SE}_{\text{lin}}$ for the regressors `PercVacant`, `PercCommercial` and `PercIndustrial`, and $\boldsymbol{SE}_{\text{boot}}, \boldsymbol{SE}_{\text{sand}} < \boldsymbol{SE}_{\text{lin}}$ for `Intercept`, `MedianIncome ($K)`, `PercResidential`. Only for `PercMinority` is $\boldsymbol{SE}_{\text{lin}}$ off by less than 10% from $\boldsymbol{SE}_{\text{boot}}$ and $\boldsymbol{SE}_{\text{sand}}$. The discrepancies affect outcomes of some of the t-tests: Under linear models theory the regressors `PercCommercial` and `PercIndustrial` have sizable t-values of 2.700 and 2.818, respectively, which are reduced to unconvincing values below 1.9 and 1.6, respectively, if the x–y bootstrap or the sandwich estimator are used. On the other hand, for `MedianIncome ($K)` the t-value −0.977 from linear models theory becomes borderline significant with the bootstrap (−1.601) or sandwich (−1.696) estimator under a plausible one-sided alternative.

A similar exercise with fewer discrepancies but similar conclusions is shown in Appendix B for the Boston Housing data.

Conclusions: (1) $\boldsymbol{SE}_{\text{boot}}$ and $\boldsymbol{SE}_{\text{sand}}$ are in substantial agreement; (2) $\boldsymbol{SE}_{\text{lin}}$ on the one hand and $\{\boldsymbol{SE}_{\text{boot}}, \boldsymbol{SE}_{\text{sand}}\}$ on the other hand can have substantial discrepancies; (3) the discrepancies are specific to regressors.

3. THE POPULATION FRAMEWORK FOR LINEAR OLS

As noted earlier, model-robust inference needs a target of estimation that is well-defined outside the linear working model. To this end, we need notation for data

[5]The response is the raw number of homeless in a census tract. The tracts do not differ by magnitudes and, according to experts, size effects seem minor. The homeless tend to clump in certain areas within census tracts, and it is thought that the regressors describe features of the tracts that make them magnets for the homeless. Finally, policy makers are accustomed to thinking in counts, not percentages.

distributions that are free of model assumptions, essentially relying on i.i.d. sampling of x–y tuples. Subsequently, OLS parameters can be introduced as statistical functionals of these distributions through best linear approximation. This is sometimes called "projection," meaning that the assumption-free data distribution is "projected" to the "nearest" distribution in the working model.

3.1 Populations for OLS Linear Regression

In an assumption-lean, model-robust population framework for OLS linear regression with random regressors, the ingredients are *regressor random variables* $X_1, \ldots, X_p$ and a *response random variable* Y. For now, the only assumption is that they are all numeric and have a joint distribution, written as

$$\boldsymbol{P} = \boldsymbol{P}(\mathrm{d}y, \mathrm{d}x_1, \ldots, \mathrm{d}x_p).$$

Data will consist of i.i.d. multivariate samples from this joint distribution (Section 5). ***No working model for $\boldsymbol{P}$ will be assumed***.

It is convenient to add a fixed regressor 1 to accommodate an intercept parameter; we may hence write

$$\vec{\boldsymbol{X}} = (1, X_1, \ldots, X_p)'$$

for the *column random vector* of the regressor variables, and $\vec{\boldsymbol{x}} = (1, x_1, \ldots, x_p)'$ for its values. We further write

$$\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}} = \boldsymbol{P}, \qquad \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}, \qquad \boldsymbol{P}_{\vec{\boldsymbol{X}}},$$

for, respectively, the joint distribution of $(Y, \vec{\boldsymbol{X}})$, the conditional distribution of Y given $\vec{\boldsymbol{X}}$, and the marginal distribution of $\vec{\boldsymbol{X}}$. These denote *actual* data distributions, free of assumptions of a working model.

All variables will be assumed to be square integrable. Required is also that $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']$ is full-rank, but permitted are nonlinear degeneracies among regressors as when they are functions of underlying independent variables such as in polynomial or B-spline regression or product interactions.

3.2 Targets of Estimation: The OLS Statistical Functional

We write any function $f(X_1, \ldots, X_p)$ of the regressors as $f(\vec{\boldsymbol{X}})$. We will need notation for the "true response surface" $\mu(\vec{\boldsymbol{X}})$, which is the conditional expectation of Y given $\vec{\boldsymbol{X}}$, and the best $L_2(\boldsymbol{P})$ approximation to Y among functions of $\vec{\boldsymbol{X}}$. It is ***not*** assumed to be linear in $\vec{\boldsymbol{X}}$:

$$\mu(\vec{\boldsymbol{X}}) \triangleq \boldsymbol{E}[Y|\vec{\boldsymbol{X}}] = \operatorname*{argmin}_{f(\vec{\boldsymbol{X}})\in L_2(\boldsymbol{P})} \boldsymbol{E}\big[(Y - f(\vec{\boldsymbol{X}}))^2\big].$$

The main definition concerns ***the best population linear approximation*** to Y, which is the linear function $l(\vec{\boldsymbol{X}}) = \boldsymbol{\beta}'\vec{\boldsymbol{X}}$ with coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ given by

$$\begin{aligned}\boldsymbol{\beta}(\boldsymbol{P}) &\triangleq \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \boldsymbol{E}\big[(Y - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2\big] \\ &= \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}Y] \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \boldsymbol{E}\big[(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2\big] \\ &= \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})].\end{aligned}$$

Both the second and fourth expressions follow from the population normal equations:

$$\begin{aligned}&\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}Y] \\ &\quad = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})] = \boldsymbol{0}.\end{aligned} \tag{4}$$

The population coefficients $\boldsymbol{\beta}(\boldsymbol{P}) = (\beta_0(\boldsymbol{P}), \beta_1(\boldsymbol{P}), \ldots, \beta_p(\boldsymbol{P}))'$ form a ***vector statistical functional***, $\boldsymbol{P} \mapsto \boldsymbol{\beta}(\boldsymbol{P})$, defined for a large class of joint data distributions $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$. If the response surface under $\boldsymbol{P}$ happens to be linear, $\mu(\vec{\boldsymbol{X}}) = \tilde{\boldsymbol{\beta}}'\vec{\boldsymbol{X}}$, as it is for example under a Gaussian linear model, $Y|\vec{\boldsymbol{X}} \sim \mathcal{N}(\tilde{\boldsymbol{\beta}}'\vec{\boldsymbol{X}}, \sigma^2)$, then $\boldsymbol{\beta}(\boldsymbol{P}) = \tilde{\boldsymbol{\beta}}$. The statistical functional is therefore a natural extension of the traditional meaning of a model parameter, justifying the notation $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$. The point is, however, that $\boldsymbol{\beta}(\cdot)$ is defined even when linearity does not hold. (Depending on the context, we may write $\boldsymbol{\beta}$ to mean $\boldsymbol{\beta}(\boldsymbol{P})$.)

3.3 The Noise-Nonlinearity Decomposition for Population OLS

The response Y has the following canonical decompositions:

$$\begin{aligned}Y &= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \underbrace{(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\vec{\boldsymbol{X}})} + \underbrace{(Y - \mu(\vec{\boldsymbol{X}}))} \\ &= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \qquad \underbrace{\eta(\vec{\boldsymbol{X}}) \qquad + \qquad \epsilon} \\ &= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \qquad\qquad\qquad \delta.\end{aligned} \tag{5}$$

We call $\epsilon = \epsilon|\vec{\boldsymbol{X}}$ the *noise* and $\eta = \eta(\vec{\boldsymbol{X}})$ the *nonlinearity*,[6] while for δ there is no standard term, so "*population residual*" may suffice; see Table 2 and Figure 1. Important to note is that (5) is a *decomposition, not a*

[6]The term "nonlinearity" has two meanings, related to each other. "The/a nonlinearity" refers to $\eta(\vec{\boldsymbol{x}})$, but "presence of nonlinearity" is a property of $\mu(\vec{\boldsymbol{x}})$.

TABLE 2
Random variables and their canonical decompositions

$\eta(\vec{X})$	$= \mu(\vec{X}) - \boldsymbol{\beta}'\vec{X}$	$= \eta$	***nonlinearity***
ϵ	$= Y - \mu(\vec{X})$		***noise***
δ	$= Y - \boldsymbol{\beta}'\vec{X}$	$= \eta + \epsilon$	***population residual***
$\mu(\vec{X})$	$= \boldsymbol{\beta}'\vec{X} + \eta(\vec{X})$		***response surface***
Y	$= \boldsymbol{\beta}'\vec{X} + \eta(\vec{X}) + \epsilon$	$= \boldsymbol{\beta}'\vec{X} + \delta$	***response***

model assumption. In a model-robust framework, there is no notion of "error term" in the usual sense; its place is taken by the population residual δ which satisfies few of the usual assumptions made in generative models. It naturally decomposes into a systematic component, the nonlinearity $\eta = \eta(\vec{X})$, and a random component, the noise $\epsilon = \epsilon|\vec{X}$. Model-trusting linear modeling, before conditioning on $\vec{X}$, must assume $\eta(\vec{X}) \overset{P}{=} 0$ and ϵ to have the same $\vec{X}$-conditional distribution in all of regressor space, that is, to be independent of $\vec{X}$. No such assumptions are made here. What is left are orthogonalities satisfied by η and ϵ in relation to $\vec{X}$. If we call independence "strong-sense orthogonality," we have instead

$$\begin{aligned} &\text{weak-sense orthogonality:} \quad \eta \perp \vec{X} \\ &\big(\boldsymbol{E}[\eta \cdot X_j] = 0 \ \ \forall j = 0, 1, \ldots, p\big), \\ &\text{medium-sense orthogonality:} \quad \epsilon \perp L_2(\boldsymbol{P}_{\vec{X}}) \\ &\big(\boldsymbol{E}\big[\epsilon \cdot f(\vec{X})\big] = 0 \ \ \forall f \in L_2(\boldsymbol{P}_{\vec{X}})\big). \end{aligned} \tag{6}$$

These are not assumptions but consequences of population OLS and the definitions. Because of the inclusion of an intercept ($j = 0$ and $f = 1$, resp.), both the nonlinearity and noise are marginally centered: $\boldsymbol{E}[\eta] = \boldsymbol{E}[\epsilon] = 0$. Importantly, it also follows that $\epsilon \perp \eta(\vec{X})$ because η is just some $f \in L_2(\boldsymbol{P}_{\vec{X}})$.

In what follows, we will need the following natural definitions:

- ***Conditional noise variance***: The noise ϵ, not assumed homoskedastic, can have arbitrary conditional distributions $\boldsymbol{P}(d\epsilon|\vec{X} = \vec{x})$ for different $\vec{x}$ except for conditional centering and finite conditional variances. Define

$$\sigma^2(\vec{X}) \overset{\Delta}{=} \boldsymbol{V}[\epsilon|\vec{X}] = \boldsymbol{E}\big[\epsilon^2|\vec{X}\big] \overset{P}{<} \infty. \tag{7}$$

  When we use the abbreviation σ^2, we will mean $\sigma^2 = \sigma^2(\vec{X})$ as we will never assume homoskedasticity.

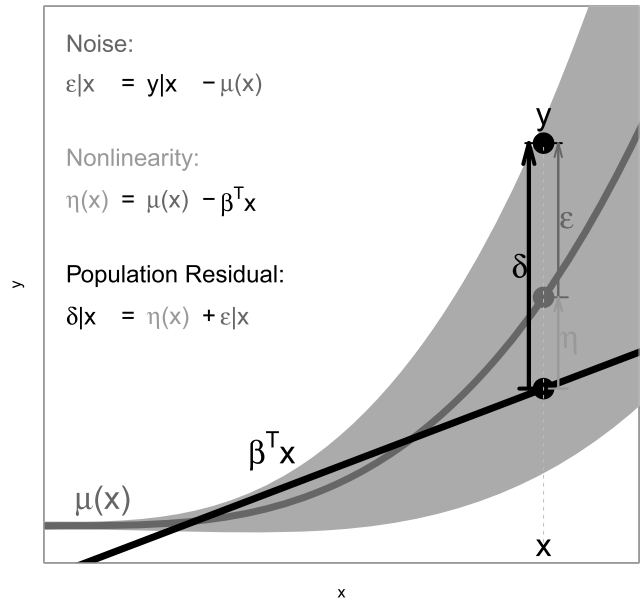

FIG. 1. *Illustration of the decomposition* (5) *for linear OLS.*

- ***Conditional mean squared error***: This is the conditional MSE of Y w.r.t. the population linear approximation $\boldsymbol{\beta}'\vec{X}$. Its definition and bias-variance decomposition are

$$m^2(\vec{X}) \overset{\Delta}{=} \boldsymbol{E}\big[\delta^2|\vec{X}\big] = \eta^2(\vec{X}) + \sigma^2(\vec{X}). \tag{8}$$

  The right-hand side follows from $\delta = \eta + \epsilon$ and $\epsilon \perp \eta(\vec{X})$ noted after (6).

In the above definitions and statements, randomness of the regressor vector $\vec{X}$ has started to play a role. The next section will discuss a crucial role of the marginal regressor distribution $\boldsymbol{P}_{\vec{X}}$.

4. BROKEN REGRESSOR ANCILLARITY I: NONLINEARITY AND RANDOM X JOINTLY AFFECT SLOPES

4.1 Misspecification Destroys Regressor Ancillarity

Conditioning on the regressors and treating them as fixed when they are random has historically been justified with the ancillarity principle. Regressor ancillarity is a property of working models $p(y|\vec{x}; \boldsymbol{\theta})$ for the conditional distribution of $Y|\vec{X}$, where $\boldsymbol{\theta}$ is the parameter of interest in the usual meaning of a parametric model. Because we treat $\vec{X}$ as random, the assumed joint distribution of $(Y, \vec{X})$ is

$$p(y, \vec{x}; \boldsymbol{\theta}) = p(y|\vec{x}; \boldsymbol{\theta})\, p(\vec{x}),$$

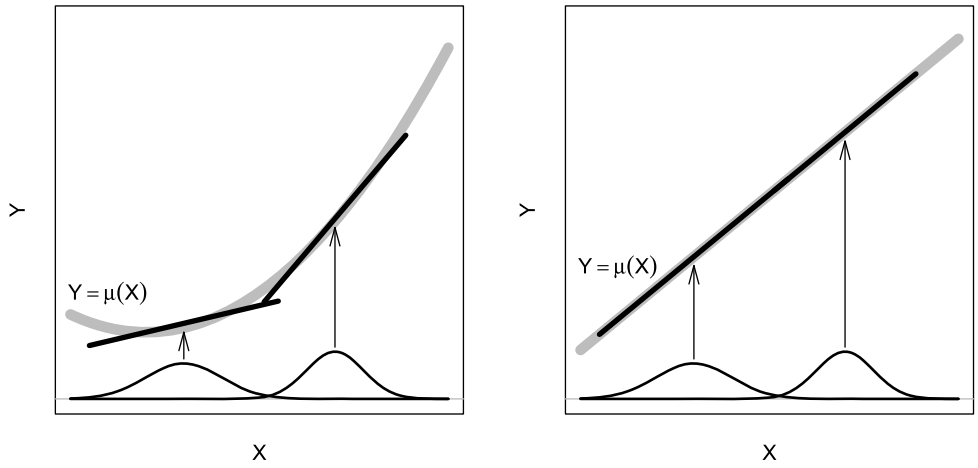

FIG. 2. *Illustration of the dependence of the population OLS solution on the marginal distribution of the regressors*: *The left figure shows dependence in the presence of nonlinearity*; *the right figure shows independence in the presence of linearity.*

where $p(\vec{x})$ is the unknown marginal regressor distribution, acting as a "nonparametric nuisance parameter." Ancillarity of $p(\vec{x})$ in relation to $\boldsymbol{\theta}$ is immediately recognized by forming likelihood ratios,

$$p(y, \vec{x}; \boldsymbol{\theta}_1)/p(y, \vec{x}; \boldsymbol{\theta}_2) = p(y|\vec{x}; \boldsymbol{\theta}_1)/p(y|\vec{x}; \boldsymbol{\theta}_2),$$

which are free of $p(\vec{x})$, detaching the regressor distribution from inference about the parameter $\boldsymbol{\theta}$. (For more on ancillarity, see Appendix C.) This logic is valid if $p(y|\vec{x}; \boldsymbol{\theta})$ correctly describes the actual conditional regressor distribution $\boldsymbol{P}_{Y|\vec{X}}$ for some $\boldsymbol{\theta}$. If this is not so, the ancillarity argument does not apply.

To pursue the consequences of nonancillarity, one needs to consider $\boldsymbol{P}_{Y|\vec{X}}$ not in the working model and interpret parameters as statistical functionals:

PROPOSITION 4.1 (Breaking regressor ancillarity in linear OLS). *Consider joint distributions that share a function $\mu(\vec{x})$ as a (a.s.) version of their conditional expectation of the response. Among these distributions, there exist $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ with $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$ if and only if $\mu(\vec{x})$ is nonlinear.*

See Appendix E.1. Because $\boldsymbol{\beta}(\boldsymbol{P}_{1,2})$ depend on Y only through $\mu(\vec{X})$, the cause of $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$ must be a difference in their regressor distributions.

The proposition is best explained graphically: Figure 2 shows single regressor scenarios with nonlinear and linear mean functions, respectively, and the same two regressor distributions. The two population OLS lines for the two regressor distributions differ in the nonlinear case and they are identical in the linear case.[7]

[7] See also White (1980a, p. 155f); his $g(Z) + \epsilon$ is our Y.

Ancillarity of regressors is sometimes informally explained as the regressor distribution being independent of, or unaffected by, the parameters of interest. From the present point of view where parameters are not labels for distributions but rather statistical functionals, this phrasing has things upside down:

It is not the parameters that affect the regressor distribution;
it is the regressor distribution that affects the parameters.

4.2 Implications of the Dependence of Slopes on Regressor Distributions

A first practical implication, illustrated by Figure 2, is that two empirical studies that use the same regressors, the same response, and the same model, may yet estimate different parameter values, $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$. This possibility arises even if the true response surface $\mu(\vec{x})$ is identical between the studies. The reason is model misspecification and differences between the regressor distributions in the two studies. Here is therefore a potential cause of so-called "parameter heterogeneity" in meta-analyses. The single-regressor situation of Figure 2 gives only an insufficient impression of the problem because for a single regressor such differences between regressor distributions are easily detected. For multiple regressors, the differences take on a multivariate nature and may become undetectable.

A second practical implication, illustrated by Figure 3, is that misspecification is a function of the regressor range: Over a narrow range a model has a better

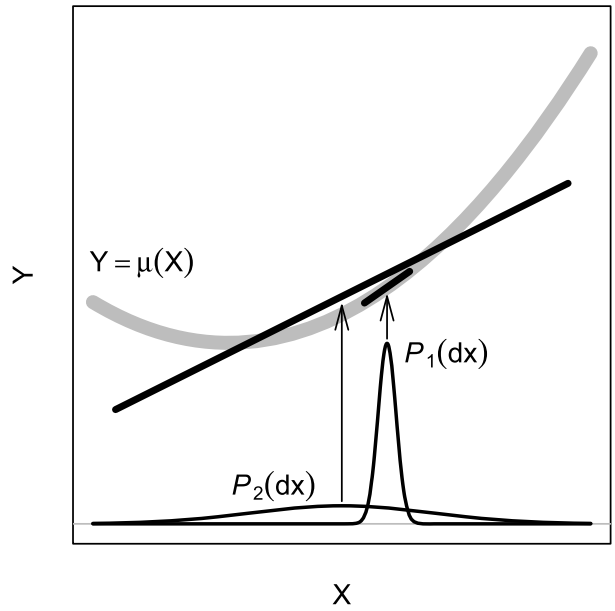

FIG. 3. *Illustration of the interplay between regressors' high–density range and nonlinearity*: *Over the small range of* $\boldsymbol{P}_1$ *the nonlinearity is undetectable and immaterial for realistic sample sizes*, *whereas over the extended range of* $\boldsymbol{P}_2$ *the nonlinearity is more likely to be detectable and relevant*.

chance of appearing "correctly specified." In the figure, over the narrow range of $\boldsymbol{P}_1(d\vec{\boldsymbol{x}})$ the linear approximation appears nearly correctly specified, whereas over the wide range of $\boldsymbol{P}_2(d\vec{\boldsymbol{x}})$ it is grossly misspecified. Again, the issue gets magnified for larger numbers of regressors where the notion of "regressor range" has a multivariate meaning.

Finally, the fact that all models have limited ranges of "acceptable approximation" is a universal issue. This holds even in physical sciences based on the most successful theories known to us.

5. THE NOISE-NONLINEARITY DECOMPOSITION OF OLS ESTIMATES

We turn to estimation from i.i.d. data.[8] We denote i.i.d. observations from a joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ by $(Y_i, \vec{\boldsymbol{X}}_i') = (Y_i, 1, X_{i,1}, \ldots, X_{i,p})$ $(i = 1, 2, \ldots, N)$. We stack them to vectors and matrices as in Table 3, inserting a constant 1 in the regressors to accommodate an intercept term. In particular, $\vec{\boldsymbol{X}}_i'$ is the ith row and $\mathbf{X}_j$ the jth column of the regressor matrix $\mathbf{X}$ $(i = 1, \ldots, N,\ j = 0, \ldots, p)$.

[8] In econometrics, where misspecification has been an important topic, the assumption of i.i.d. data is too limiting; instead, one assumes time series structures. See, for example, White (1994).

The (unknown) nonlinearity η, noise ϵ and population residual δ generate random N-vectors when evaluated at all N observations (again, see Table 3):

$$\begin{aligned} \boldsymbol{\eta} &= \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}, \\ \boldsymbol{\delta} &= \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\eta} + \boldsymbol{\epsilon}. \end{aligned} \tag{9}$$

It is important to distinguish between population and sample properties: The vectors $\boldsymbol{\delta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are *not* orthogonal to the regressor columns $\mathbf{X}_j$ in the sample. Writing $\langle \cdot, \cdot \rangle$ for the usual Euclidean inner product on $\mathbb{R}^N$, we have in general

$$\langle \boldsymbol{\delta}, \mathbf{X}_j \rangle \neq 0, \qquad \langle \boldsymbol{\epsilon}, \mathbf{X}_j \rangle \neq 0, \qquad \langle \boldsymbol{\eta}, \mathbf{X}_j \rangle \neq 0,$$

even though the associated random variables are orthogonal to X_j in the population: $\boldsymbol{E}[\delta X_j] = 0$, $\boldsymbol{E}[\epsilon X_j] = 0$, $\boldsymbol{E}[\eta(\vec{\boldsymbol{X}})X_j] = 0$, according to (6).

The *OLS estimate* of $\boldsymbol{\beta}(\boldsymbol{P})$ is as usual

$$\boxed{\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\tilde{\boldsymbol{\beta}}} \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\|^2 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.} \tag{10}$$

If we write $\hat{\boldsymbol{P}}$ for the empirical distribution of the observations $(Y_i, \vec{\boldsymbol{X}}_i')$, then $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$ is the plug-in estimate. Associated is the sample residual vector $\boldsymbol{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, based on $\hat{\boldsymbol{\beta}}$, which is distinct from the population residual vector $\boldsymbol{\delta} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, based on $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$.

In linear models theory which conditions on (or fixes) $\mathbf{X}$, the target of estimation is what we may call the "$\mathbf{X}$-*conditional parameter*":

$$\boxed{\begin{aligned} \boldsymbol{\beta}(\mathbf{X}) &\triangleq \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] \\ &= \operatorname{argmin}_{\boldsymbol{\beta}} \boldsymbol{E}\big[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2|\mathbf{X}\big] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}. \end{aligned}} \tag{11}$$

In random-$\mathbf{X}$ theory, on the other hand, the target of estimation is $\boldsymbol{\beta}(\boldsymbol{P})$, while the $\mathbf{X}$-conditional parameter $\boldsymbol{\beta}(\mathbf{X})$ is a random vector. The vectors $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$, $\boldsymbol{\beta}(\mathbf{X})$ and $\boldsymbol{\beta}(\boldsymbol{P})$ lend themselves to the following telescoping decomposition:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\boldsymbol{P}) = \big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\mathbf{X})\big) + \big(\boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P})\big), \tag{12}$$

which in turn reflects the decomposition $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$:

DEFINITION AND LEMMA. Define "Estimation Offsets" (EOs) as follows:

$$\boxed{\begin{aligned} \textit{Total EO} &\triangleq \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\boldsymbol{P}) &&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\delta}, \\ \textit{Noise EO} &\triangleq \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\mathbf{X}) &&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}, \\ \textit{Approx. EO} &\triangleq \boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P}) &&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta}. \end{aligned}} \tag{13}$$

TABLE 3
Random variable notation for estimation in linear OLS based on i.i.d. observational data

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$,		parameter vector	$((p+1)\times 1)$
$\mathbf{Y} = (Y_1, \ldots, Y_N)'$,		response vector	$(N \times 1)$
$\mathbf{X}_j = (X_{1,j}, \ldots, X_{N,j})'$,		jth regressor vector	$(N \times 1)$
$\mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_p]$	$= \begin{bmatrix} \vec{X}_1' \\ \cdots \\ \cdots \\ \vec{X}_N' \end{bmatrix}$,	regressor matrix with intercept	$(N \times (p+1))$
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)'$,	$\mu_i = \mu(\vec{X}_i) = \boldsymbol{E}[Y\|\vec{X}_i]$,	conditional means	$(N \times 1)$
$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)'$,	$\eta_i = \eta(\vec{X}_i) = \mu_i - \boldsymbol{\beta}'\vec{X}_i$,	nonlinearities	$(N \times 1)$
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)'$,	$\epsilon_i = Y_i - \mu_i$,	noise values	$(N \times 1)$
$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)'$,	$\delta_i = \eta_i + \epsilon_i$,	population residuals	$(N \times 1)$
$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)'$,	$\sigma_i = \sigma(\vec{X}_i) = \boldsymbol{V}[Y\|\vec{X}_i]^{1/2}$,	conditional sdevs	$(N \times 1)$
$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$	$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$,	parameter estimates	$((p+1)\times 1)$
$\boldsymbol{r} = (r_1, \ldots, r_N)'$	$= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$,	sample residuals	$(N \times 1)$

The right-hand sides follow from (9): $\boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}$, $\boldsymbol{\delta} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

$$\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu},$$

$$\boldsymbol{\beta}(\boldsymbol{P}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}).$$

The first defines $\hat{\boldsymbol{\beta}}$, the second uses $\boldsymbol{E}[\mathbf{Y}|\mathbf{X}] = \boldsymbol{\mu}$, and the third is a tautology.

REMARK. One might be tempted to interpret the approximation EO $\boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P})$ as a bias because it is the difference of two targets of estimation. This interpretation is entirely wrong. The approximation EO is a random variable when nonlinearity is present. It will be seen to contribute not a bias but a $N^{-1/2}$ order term to the sampling variability of $\hat{\boldsymbol{\beta}}$ (Section 7).

6. BROKEN REGRESSOR ANCILLARITY II: NONLINEARITY AND RANDOM X CREATE SAMPLING VARIATION

6.1 Sampling Variation's Two Sources: Noise and Nonlinearity

For the $\mathbf{X}$-conditional parameter $\boldsymbol{\beta}(\mathbf{X})$ to be a nontrivial random variable, two factors need to be present: (1) the regressors $\vec{X}$ need to be random and (2) the nonlinearity $\eta(\vec{X})$ must not vanish: $\boldsymbol{P}[\eta(\vec{X}) \neq 0] > 0$. These factors conspire to produce sampling variation according to (13), which shows the approximation EO to depend on the random matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the vector of nonlinearity values $\boldsymbol{\eta}$. We have

$$\boldsymbol{V}[\hat{\boldsymbol{\beta}}] = \boldsymbol{E}\big[\boldsymbol{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\big] + \boldsymbol{V}\big[\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\big], \tag{14}$$

where the left-hand side represents the full unconditional variability of $\hat{\boldsymbol{\beta}}$ relevant for statistical inference. By the definition of EOs and the lemma in Section 5 this decomposition parallels $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$:

$$\boldsymbol{V}[\hat{\boldsymbol{\beta}}] = \boldsymbol{V}\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\delta}\big],$$

$$\boldsymbol{E}\big[\boldsymbol{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\big] = \boldsymbol{E}\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{V}[\boldsymbol{\epsilon}|\mathbf{X}]\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\big],$$

$$\boxed{\boldsymbol{V}\big[\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\big] = \boldsymbol{V}\big[\boldsymbol{\beta}(\mathbf{X})\big] = \boldsymbol{V}\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta}\big]}$$

The center line represents the marginal sampling variability due to noise combined with randomness in $\mathbf{X}$. Note that $\boldsymbol{V}[\boldsymbol{\epsilon}|\mathbf{X}] = D_{\boldsymbol{\sigma}^2}$ is the diagonal matrix of noise variances. The box shows how the vector of nonlinearities $\boldsymbol{\eta}$ "conspires" with the randomness of $\mathbf{X}$ to generate sampling variability in $\boldsymbol{\beta}(\mathbf{X})$.

Intuition for the sampling variability of $\boldsymbol{\beta}(\mathbf{X})$ is best provided by a graphical illustration. In order to isolate this effect, we consider a noise-free situation where the response is deterministic and nonlinear, hence a linear fit is "misspecified." To this end, let $Y = \mu(\vec{X})$ where $\mu(\cdot)$ is some nonlinear function, and hence $\boldsymbol{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \mathbf{0}$ a.s. An example is shown in the left-hand frame of Figure 4 for a single regressor, with OLS lines fitted to two "datasets" consisting of $N = 5$ regressor values each. The randomness in the regressors causes the fitted line to differ between datasets, hence exhibit sampling variability due to the nonlinearity of the response. This effect is absent in the right-hand frame of Figure 4 where the response is linear.[9]

[9]We remind the reader of the more striking animated illustration of this effect by executing the line of code shown in footnote 3.

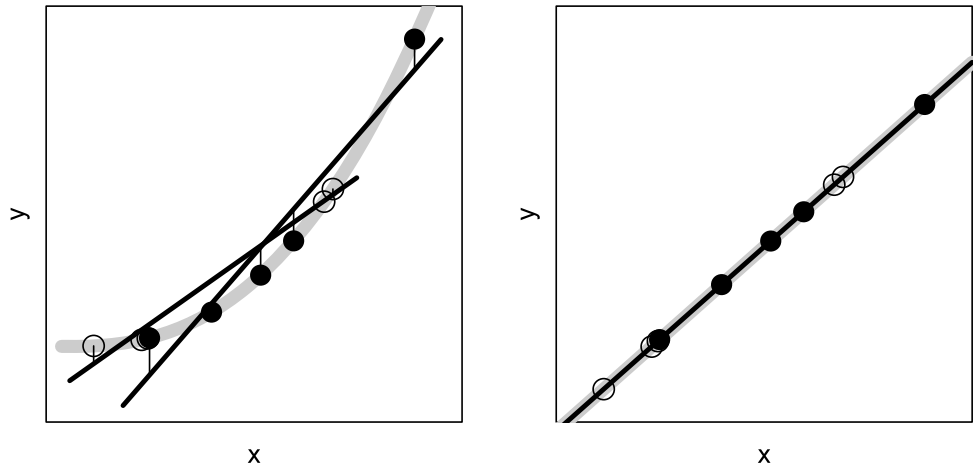

FIG. 4. *Noise-less response*: *The filled and the open circles represent two "datasets" from the same population. The x-values are random*; *the y-values are a deterministic function of x*: $y = \mu(x)$ (*shown in gray*). *Left*: *The true response $\mu(x)$ is nonlinear*; *the open and the filled circles have different OLS lines* (*shown in black*). *Right*: *The true response $\mu(x)$ is linear*; *the open and the filled circles have the same OLS line* (*black on top of gray*).

6.2 Quandaries of Fixed-X Theory and the Need for Random-X Theory

The fixed-$\mathbf{X}$ approach of linear models theory necessarily assumes correct specification. Its only source of sampling variability is the noise EO $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\mathbf{X})$ arising from the conditional response distribution, ignoring the approximation EO $\boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P})$ due to conditioning on $\mathbf{X}$. A partial remedy in fixed-$\mathbf{X}$ theory is to rely on diagnostics to detect lack of fit (misspecification). We emphasize that diagnostics should be part of every regression analysis. In fact, to assist such diagnostics and make them relevant for correctly sized standard errors, we propose in Section 12 a test to identify slopes that may have their usual standard errors invalidated by misspecification. Furthermore, in Part II we propose a misspecification diagnostic for regression parameters.

Data analysts may not stop with negative findings from model diagnostics and instead continue with data-driven model improvement by, for example, transforming variables and adding terms to the fitted equation till the residuals "look right." However, model improvement based on the data can have drawbacks and limits. A drawback is that it can invalidate subsequent inferences in unpredictable ways, as does any data-driven variable selection, formal or informal (see, e.g., Berk et al., 2013, Lee et al., 2016). A limitation is that residual diagnostics lose power as the number of regressors increases. This fact follows from what we may call "Mammen's dilemma." Mammen (1996) showed, roughly speaking, that for models with numerous regressors the residual distribution tends to look as assumed by the working model, for example, Gaussian for OLS, Laplacian for LAD, irrespective of the true error distribution. For these reasons, data analysts who diagnose and improve their models will find themselves torn at some point between hunches of having done too much of a good thing and missing out on something.

In light of such uncertainties arising from diagnostics and model improvement, it may be of some comfort that tools are available for asymptotically correct inference under model misspecification, including misspecified deterministic responses ($Y = \mu(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}}) = 0$). These tools—sandwich and x–y bootstrap[10] estimators of standard error—derive their justification from central limit theorems (CLTs) to be described next.

7. MODEL-ROBUST CLTS, CANONICALLY DECOMPOSED

Random-$\mathbf{X}$ CLTs for OLS are standard, and the novel aspect of the following proposition is in decomposing the overall asymptotic variance into contributions stemming from the noise EO and the approximation EO according to (13), thereby providing an asymp-

[10]It needs to pointed out again that the residual bootstrap is not assumption-lean. It requires the population residual δ to be a conventional error term, i.i.d. across the N observations, implying first and second order correct specification ($\eta(\vec{\boldsymbol{X}}) = 0$ and $\sigma^2(\vec{\boldsymbol{X}}) = \sigma^2$ constant). The only lean aspect is that the error term no longer needs to be Gaussian.

totic analog of the finite-sample decomposition (14) of sampling variance in Section 6.1.

PROPOSITION 7.1. *For linear OLS, the three EOs follow CLTs*:

$$\boxed{\begin{aligned}&\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \\ &\quad \mathcal{N}(\mathbf{0}, \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\,\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}) \\ &\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}(\mathbf{X})) \xrightarrow{\mathcal{D}} \\ &\quad \mathcal{N}(\mathbf{0}, \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\,\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}) \\ &\sqrt{N}(\boldsymbol{\beta}(\mathbf{X})-\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \\ &\quad \mathcal{N}(\mathbf{0}, \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\,\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1})\end{aligned}}$$

These statements again reflect the decomposition (8), $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$. According to (7) and (8), $m^2(\vec{\boldsymbol{X}})$ can be replaced by δ^2 and $\sigma^2(\vec{\boldsymbol{X}})$ by ϵ^2:

$$\begin{aligned}\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] &= \boldsymbol{E}[\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'], \\ \boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] &= \boldsymbol{E}[\epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'].\end{aligned} \tag{15}$$

The asymptotic variance of linear OLS can therefore be written as

$$\boldsymbol{AV}[\boldsymbol{P};\boldsymbol{\beta}] \triangleq \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}. \tag{16}$$

As always, $\boldsymbol{\beta}$ stands for the statistical functional $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ and by implication its plug-in OLS estimator $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$. The formula is the basis for plug-in that produces the sandwich estimator of standard error (Section 8.1).

Special cases covered by the above proposition are the following:

- *First-order correct specification:* $\eta(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$. The sandwich form is solely due to heteroskedasticity.
- *Deterministic nonlinear response:* $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$. The sandwich form is solely due to the nonlinearity and randomness of $\mathbf{X}$.
- *First- and second-order correct specification:* $\eta(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$, $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} \sigma_0^2$. The *non*sandwich form is asymptotically valid without Gaussianity: $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma_0^2\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1})$.

8. SANDWICH ESTIMATORS AND THE M-OF-N BOOTSTRAP

Empirically one observes that standard error estimates obtained from the x–y bootstrap and from the sandwich estimator are generally close to each other (Section 2). This is intuitively unsurprising as they both estimate the same asymptotic variance, that of the first CLT in Proposition 7.1. A closer connection between them will be described here and established in generality in Part II.[11]

8.1 The Plug-in Sandwich Estimator of Asymptotic Variance

Plug-in estimators of standard error are obtained by substituting the empirical distribution $\hat{\boldsymbol{P}}$ for the true $\boldsymbol{P}$ in formulas for asymptotic variances. As the asymptotic variance $\boldsymbol{AV}[\boldsymbol{P};\boldsymbol{\beta}]$ in (16) is given explicitly and is suitably continuous in the two arguments, one obtains a consistent estimator by plugging in $\hat{\boldsymbol{P}}$ for $\boldsymbol{P}$:

$$\begin{aligned}\hat{\boldsymbol{AV}}[\boldsymbol{\beta}] &\triangleq \boldsymbol{AV}[\boldsymbol{\beta}, \hat{\boldsymbol{P}}], \\ \hat{\boldsymbol{SE}}[\beta_j] &\triangleq \frac{1}{N^{1/2}}(\hat{\boldsymbol{AV}}[\boldsymbol{\beta}])_{jj}^{1/2}.\end{aligned} \tag{17}$$

[Recall again that $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ stands for the OLS statistical functional which specializes to its plug-in estimator through $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$.] Concretely, one estimates expectations $\boldsymbol{E}[\ldots]$ with sample means $\hat{\boldsymbol{E}}[\ldots]$, $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ with $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$, and hence population residuals $\delta^2 = (Y - \vec{\boldsymbol{X}}\boldsymbol{\beta})^2$ with sample residuals $r_i^2 = (Y_i - \vec{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})^2$. Collecting the latter in a diagonal matrix $\boldsymbol{D}_{\boldsymbol{r}}^2$, one has

$$\begin{aligned}\hat{\boldsymbol{E}}[r^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] &= \frac{1}{N}(\mathbf{X}'\boldsymbol{D}_{\boldsymbol{r}}^2\mathbf{X}), \\ \hat{\boldsymbol{E}}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] &= \frac{1}{N}(\mathbf{X}'\mathbf{X}).\end{aligned}$$

The sandwich estimator $\hat{\boldsymbol{AV}}_{\text{sand}}[\boldsymbol{\beta}] = \hat{\boldsymbol{AV}}[\boldsymbol{\beta}]$ for linear OLS in its original form (White, 1980a) is therefore obtained explicitly as follows:

$$\begin{aligned}\hat{\boldsymbol{AV}}_{\text{sand}}[\boldsymbol{\beta}] &\triangleq \hat{\boldsymbol{E}}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\hat{\boldsymbol{E}}[r^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\hat{\boldsymbol{E}}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1} \\ &= N(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{D}_{\boldsymbol{r}}^2\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\end{aligned} \tag{18}$$

This is version "HC" in MacKinnon and White (1985). A modification accounts for the fact that residuals have smaller variance than noise, calling for a correction by replacing $1/N^{1/2}$ in (17) with $1/(N-p-1)^{1/2}$, in analogy to the linear models estimator ("HC1" ibid.). Another modification is to correct individual residuals for their reduced variance according to $\boldsymbol{V}[r_i|\mathbf{X}] = \sigma^2(1 - H_{ii})$ under homoskedasticity and ignoring nonlinearity ("HC2" ibid.). Further modifications include a version based on the jackknife ("HC3" ibid.) using

[11] A third assumption-lean method of inference is empirical likelihood. See Owen (2001).

leave-one-out residuals. MacKinnon and White (1985) also mention that some forms of sandwich estimators were independently derived by Efron (1982, p. 18f) using the infinitesimal jackknife, and by Hinkley (1977) using a "weighted jackknife." See Weber (1986) for a concise comparison in the linear model limited to heteroskedasticity.

8.2 Sandwich Estimators Are Limits of M-of-N Bootstrap Estimators

An alternative to plug-in is estimating asymptotic variance with the x–y bootstrap whose justification essentially derives from the validity of the CLT. Conventionally the resample size, here denoted by M, is taken to be the same as the sample size N, but it is useful to distinguish between these two quantities and allow the resample size M to differ from N, resulting in the "M-of-N bootstrap." One distinguishes

- M-of-N bootstrap resampling *with* replacement from
- M-out-of-N subsampling *without* replacement.

In resampling, M can be any $M < \infty$; in subsampling, M must satisfy $M < N$.[12] To fix notation, denote bootstrap estimates by $\boldsymbol{\beta}^*_M = \boldsymbol{\beta}(\boldsymbol{P}^*_M)$, where $\boldsymbol{P}^*_M$ is the empirical distribution of bootstrap data $\{(Y^*_i, \vec{\boldsymbol{X}}^{*\prime}_i)\}_{i=1,\ldots,M}$ drawn i.i.d. from $\hat{\boldsymbol{P}}_N$. Bootstrap estimates of asymptotic variance are therefore

$$\hat{\boldsymbol{A}}\hat{\boldsymbol{V}}_{\text{boot}}[\boldsymbol{\beta}] \triangleq M\boldsymbol{V}_{\hat{\boldsymbol{P}}_N}[\boldsymbol{\beta}^*_M]. \tag{19}$$

The connection between bootstrap and sandwich estimates is as follows.

PROPOSITION 8.1. *The sandwich estimator* (18) *is the M-of-N bootstrap estimator* (19) *in the limit $M \to \infty$ for a fixed dataset $\hat{\boldsymbol{P}}_N$ of size N.*

See Part II for full generality. Bootstrap approaches may be more flexible than sandwich approaches because the bootstrap distribution can be used to generate confidence intervals that are second-order correct (see, e.g., Efron and Tibshirani, 1993; Hall 1992; McCarthy, Zhang et al., 2018).

[12]The M-of-N bootstrap for $M \ll N$ "works" more often than the conventional N-of-N bootstrap; see Bickel, Götze and van Zwet (1997) who showed that the favorable properties of $M \ll N$ subsampling obtained by Politis and Romano (1994) carry over to the $M \ll N$ bootstrap.

9. ADJUSTED REGRESSORS

This section prepares the ground for two projects: (1) proposing meanings of slopes in the presence of nonlinearity (Section 10), and (2) comparing standard errors of slopes, model-robust versus model-trusting (Section 11). The first requires the well-known adjustment formula for slopes in multiple regression, while the second requires adjustment formulas for standard errors, both model-trusting and model-robust. Although the adjustment formulas are standard, they will be stated explicitly to fix notation. [See Appendix D for more notational details.]

- *Adjustment in populations:* The *population-adjusted regressor random variable* $X_{j\bullet}$ is the "residual" of the population regression of X_j, used as the response, on all other regressors. The response Y can be adjusted similarly, and we may denote it by $Y_{\bullet -j}$ to indicate that X_j is not among the adjustors, which is implicit in the adjustment of X_j. The multiple regression coefficient $\beta_j = \beta_j(\boldsymbol{P})$ of the population regression of Y on $\vec{\boldsymbol{X}}$ is obtained as the simple regression through the origin of $Y_{\bullet -j}$ or Y on $X_{j\bullet}$:

$$\begin{aligned}\beta_j &= \frac{\boldsymbol{E}[Y_{\bullet -j}X_{j\bullet}]}{\boldsymbol{E}[X^2_{j\bullet}]} = \frac{\boldsymbol{E}[YX_{j\bullet}]}{\boldsymbol{E}[X^2_{j\bullet}]} \\ &= \frac{\boldsymbol{E}[\mu(\vec{\boldsymbol{X}})X_{j\bullet}]}{\boldsymbol{E}[X^2_{j\bullet}]}.\end{aligned} \tag{20}$$

  The last representation holds because $X_{j\bullet}$ is a function of $\vec{\boldsymbol{X}}$ only which permits conditioning of Y on $\vec{\boldsymbol{X}}$ in the numerator.
- *Adjustment in samples:* Define the *sample-adjusted regressor column* $\mathbf{X}_{j\hat{\bullet}}$ to be the residual vector of the sample regression of $\mathbf{X}_j$, used as the response vector, on all other regressor vectors. The response vector $\mathbf{Y}$ can be sample-adjusted similarly, and we may denote it by $\mathbf{Y}_{\hat{\bullet}-j}$ to indicate that $\mathbf{X}_j$ is not among the adjustors, which is implicit for $\mathbf{X}_{j\hat{\bullet}}$. (Note the use of hat notation "$\hat{\bullet}$" to distinguish it from population-based adjustment "$\bullet$.") The coefficient estimate $\hat{\beta}_j$ of the multiple regression of $\mathbf{Y}$ on $\mathbf{X}$ is obtained as the simple regression through the origin of $\mathbf{Y}_{\hat{\bullet}-j}$ or $\mathbf{Y}$ on $\mathbf{X}_{j\bullet}$:

$$\hat{\beta}_j = \frac{\langle \mathbf{Y}_{\hat{\bullet}-j}, \mathbf{X}_{j\hat{\bullet}}\rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^2} = \frac{\langle \mathbf{Y}, \mathbf{X}_{j\hat{\bullet}}\rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^2}. \tag{21}$$

[For practice, the patient reader may wrap his/her mind around the distinction between $\mathbf{X}_{j\hat{\bullet}}$ and $\mathbf{X}_{j\bullet}$, the latter being the vector of population-adjusted $X_{i,j\bullet}$. The components of the former are dependent, those of the latter independent.]

10. MEANINGS OF SLOPES IN THE PRESENCE OF NONLINEARITY

A first use of regressor adjustment is for proposing meanings of linear slopes in the presence of nonlinearity, and responding to Freedman's (2006, p. 302) objection: "...it is quite another thing to ignore bias [nonlinearity]. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter." Against this view, one may argue that "flawed" models are a fact of life. Flaws such as nonlinearity can go undetected, or they can be tolerated for insightful simplification. A "parameter" based on best approximation is then not intrinsically wrong but in need of a useful interpretation.

The issue is that, in the presence of nonlinearity, slopes lose their usual interpretation: β_j is no longer the average difference in Y associated with a unit difference in X_j at fixed levels of all other X_k. Such an interpretation holds for the best approximation $\boldsymbol{\beta}'\vec{\boldsymbol{x}}$ but not the conditional mean function $\mu(\vec{\boldsymbol{x}})$. The challenge is to provide an alternative interpretation that remains valid and intuitive. As mentioned, a plausible approach is to use adjusted variables, hence by (20) and (21) it is sufficient to solve the interpretation problem for simple regression through the origin. In a sense to be made precise, slopes can then be interpreted as weighted averages of "casewise" and "pairwise" slopes. To lighten the notational burden, we drop subscripts from adjusted variables:

$$\begin{aligned} y &\leftarrow Y_{\bullet -j}, \qquad x \leftarrow X_{j\bullet}, \\ \beta &\leftarrow \beta_j \quad \text{for populations,} \\ y_i &\leftarrow (\mathbf{Y}_{\hat{\bullet} -j})_i, \qquad x_i \leftarrow (\mathbf{X}_{j\hat{\bullet}})_i, \\ \hat{\beta} &\leftarrow \hat{\beta}_j \quad \text{for samples.} \end{aligned}$$

By (20) and (21), the population slopes and their estimates are, respectively,

$$\beta = \frac{E[yx]}{E[x^2]} \quad \text{and} \quad \hat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2}.$$

Slope interpretation will be based on the following devices:

- *Population parameters* β can be represented as weighted averages of ...
  - *casewise slopes*: For a random case (x, y), we have

$$\beta = \boldsymbol{E}[wb], \quad \text{where } b \triangleq \frac{y}{x},\ w \triangleq \frac{x^2}{\boldsymbol{E}[x^2]}.$$

    Thus b is the casewise slope through the origin and w its weight.
  - *pairwise slopes*: For i.i.d. cases (x, y) and (x', y'). we have

$$\beta = \boldsymbol{E}[wb],$$

$$\text{where } b \triangleq \frac{y - y'}{x - x'},\ w \triangleq \frac{(x - x')^2}{\boldsymbol{E}[(x - x')^2]}.$$

    Thus b is the pairwise slope and w its weight.
- *Sample estimates* $\hat{\beta}$ can be represented as weighted averages of ...
  - *casewise slopes*:

$$\hat{\beta} = \sum_i w_i b_i, \quad \text{where } b_i \triangleq \frac{y_i}{x_i},\ w_i \triangleq \frac{x_i^2}{\sum_{i'} x_{i'}^2}.$$

    Thus b_i are casewise slopes and w_i their weights.
  - *pairwise slopes*:

$$\hat{\beta} = \sum_{ik} w_{ik} b_{ik},$$

$$\text{where } b_{ik} \triangleq \frac{y_i - y_k}{x_i - x_k},\ w_{ik} \triangleq \frac{(x_i - x_k)^2}{\sum_{i'k'} (x_{i'} - x_{k'})^2}.$$

    Thus b_{ik} are pairwise slopes and w_{ik} their weights $(i \neq k)$.

See Figure 5 for an illustration for samples. The formulas support the intuition that, even in the presence of nonlinearity, a linear fit can describe the overall direction of the association between the response and a regressor after adjustment.

There exist of course examples where no global direction of association exists, as when $\boldsymbol{E}[y|x] \sim x^2$ and the regressor distribution $\boldsymbol{P}_x$ is symmetric about 0. There exists, however, local association, which is negative for $x < 0$ and positive for $x > 0$. If $\boldsymbol{E}[x]/\boldsymbol{SD}[x] \gg 0$, the overall direction of association is positive and a linear fit provides an excellent approximation to x^2, illustrating once again the crucial role of $\boldsymbol{P}_x$.

We conclude with a note on the history of the above formulas: Stigler (2001) points to Edgeworth, while Berman (1988) traces them back to an 1841 article by Jacobi written in Latin. A generalization based on tuples rather than pairs of cases was used by Wu (1986) for the analysis of jackknife and bootstrap procedures (see his Section 3, Theorem 1). Gelman and Park (2009) also refer to the representation of OLS slopes as weighted means of pairwise slopes.

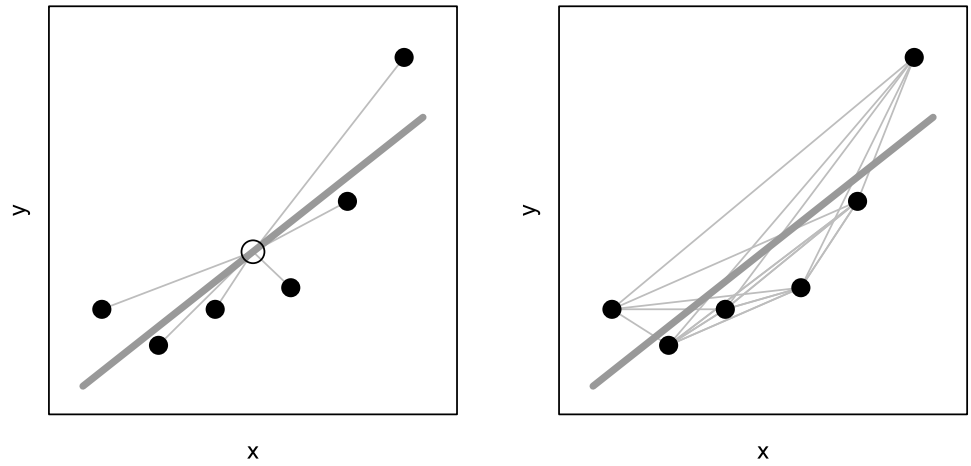

FIG. 5. *Casewise and pairwise average weighted slopes illustrated*: *Both plots show the same six points* (*"cases"*) *as well as the OLS line fitted to them* (*fat gray*). *The left-hand plot shows the casewise slopes from the mean point* (*open circle*) *to the six cases*, *while the right-hand plot shows the pairwise slopes between all* 15 *pairs*. *In both plots*, *the observed slopes are positive with just one exception each*, *supporting the impression that the direction of association is positive*.

11. ASYMPTOTIC VARIANCES—PROPER AND IMPROPER

The following prepares the ground for an asymptotic comparison of model-robust and model-trusting standard errors, one regressor at a time.

11.1 Preliminaries: Adjustment Formulas for EOs and Their CLTs

The vectorized formulas for estimation offsets (12) can be written componentwise using adjustment:

$$\begin{aligned}
&\textit{Total EO}: && \hat\beta_j - \beta_j = \frac{\langle \mathbf{X}_{j\hat\bullet}, \boldsymbol{\delta}\rangle}{\|\mathbf{X}_{j\hat\bullet}\|^2},\\
&\textit{Noise EO}: && \hat\beta_j - \beta_j(\mathbf{X}) = \frac{\langle \mathbf{X}_{j\hat\bullet}, \boldsymbol{\epsilon}\rangle}{\|\mathbf{X}_{j\hat\bullet}\|^2},\\
&\textit{Approximation EO}: && \beta_j(\mathbf{X}) - \beta_j = \frac{\langle \mathbf{X}_{j\hat\bullet}, \boldsymbol{\eta}\rangle}{\|\mathbf{X}_{j\hat\bullet}\|^2}.
\end{aligned}$$

To see these identities directly, note the following, in addition to (21): $\boldsymbol{E}[\hat\beta_j|\mathbf{X}] = \langle \boldsymbol{\mu}, \mathbf{X}_{j\hat\bullet}\rangle/\|\mathbf{X}_{j\hat\bullet}\|^2$ and $\beta_j = \langle \mathbf{X}\boldsymbol{\beta}, \mathbf{X}_{j\hat\bullet}\rangle/\|\mathbf{X}_{j\hat\bullet}\|^2$, the latter due to $\langle \mathbf{X}_{j\hat\bullet}, \mathbf{X}_k\rangle = \delta_{jk}\|\mathbf{X}_{j\hat\bullet}\|^2$. Finally use $\boldsymbol{\delta} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}$.

From the above expressions for the EOs one obtains asymptotic normality for each using population adjustment:

COROLLARY.

$$\begin{aligned}
&N^{1/2}(\hat\beta_j - \beta_j)\\
&\quad\xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})X_{j\bullet}^2]}{\boldsymbol{E}[X_{j\bullet}^2]^2}\Big) = \mathcal{N}\Big(0, \frac{\boldsymbol{E}[\delta^2 X_{j\bullet}^2]}{\boldsymbol{E}[X_{j\bullet}^2]^2}\Big),\\
&N^{1/2}(\hat\beta_j - \beta_j(\mathbf{X}))\\
&\quad\xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})X_{j\bullet}^2]}{\boldsymbol{E}[X_{j\bullet}^2]^2}\Big) = \mathcal{N}\Big(0, \frac{\boldsymbol{E}[\epsilon^2 X_{j\bullet}^2]}{\boldsymbol{E}[X_{j\bullet}^2]^2}\Big),\\
&N^{1/2}(\beta_j(\mathbf{X}) - \beta_j)\\
&\quad\xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})X_{j\bullet}^2]}{\boldsymbol{E}[X_{j\bullet}^2]^2}\Big)
\end{aligned}$$

The equalities on the right-hand side in the first and second case are based on (15). The first CLT in its right-side form is useful for plug-in estimation of asymptotic variance, one slope at a time. The sandwich form of matrices has been reduced to ratios where numerators correspond to the "meat" and squared denominators to the "breads."

11.2 Model-Robust Asymptotic Variances in Terms of Adjusted Regressors

The CLTs in the corollary of Section 11.1 contain three asymptotic variances of the same form with arguments $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$. We will use $m^2(\vec{\boldsymbol{X}})$ in the following definition for the overall asymptotic variance, but by substituting $\sigma^2(\vec{\boldsymbol{X}})$ or $\eta^2(\vec{\boldsymbol{X}})$ for $m^2(\vec{\boldsymbol{X}})$ one obtains terms that can be interpreted as components of the overall asymptotic variance or else

as asymptotic variances in the absence of nonlinearity or absence of noise.

DEFINITION. Proper asymptotic variance.

$$\boxed{\boldsymbol{AV}_{\text{lean}}[\beta_j; m^2] \triangleq \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})X_{j\bullet}^2]}{\boldsymbol{E}[X_{j\bullet}^2]^2}.}$$

From (8), $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$, one obtains

$$\boldsymbol{AV}_{\text{lean}}[\beta_j; m^2] = \boldsymbol{AV}_{\text{lean}}[\beta_j; \sigma^2] + \boldsymbol{AV}_{\text{lean}}[\beta_j; \eta^2].$$

The subscript "lean" refers to validity in the assumption-lean model-robust framework. This proper asymptotic variance will be compared to the potentially improper asymptotic variance of model-trusting linear models theory (Section 11.4).

11.3 Model-Trusting Asymptotic Variances in Terms of Adjusted Regressors

The goal is to provide an asymptotic limit for the usual model-trusting standard error estimate of linear models theory in the model-robust framework. To this end, we need the model-robust limit of the usual estimate of the noise variance, $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(N - p - 1)$ as $N \to \infty$:

$$\hat{\sigma}^2 \xrightarrow{\boldsymbol{P}} \boldsymbol{E}[\delta^2] = \boldsymbol{E}[m^2(\vec{\boldsymbol{X}})].$$

Thus the model-robust limit of $\hat{\sigma}^2$ is the average conditional MSE of Y, which again decomposes according to $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$.

Squared standard error estimates are, in matrix and adjustment form,

$$\hat{\boldsymbol{V}}_{\text{lin}}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1},$$
$$\hat{\boldsymbol{SE}}_{\text{lin}}^2[\hat{\beta}] = \frac{\hat{\sigma}^2}{\|\mathbf{X}_{j\hat{\bullet}}\|^2}. \tag{22}$$

Their assumption-lean scaled limits are

$$N\hat{\boldsymbol{V}}_{\text{lin}}[\hat{\boldsymbol{\beta}}] \xrightarrow{\boldsymbol{P}} \boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1},$$
$$N\hat{\boldsymbol{SE}}_{\text{lin}}^2[\hat{\beta}_j] \xrightarrow{\boldsymbol{P}} \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[X_{j\bullet}^2]}.$$

DEFINITION. Improper asymptotic variance.

$$\boxed{\boldsymbol{AV}_{\text{lin}}[\beta_j; m^2] \triangleq \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[X_{j\bullet}^2]}.}$$

This decomposes as usual:

$$\boldsymbol{AV}_{\text{lin}}[\beta_j; m^2] = \boldsymbol{AV}_{\text{lin}}[\beta_j; \sigma^2] + \boldsymbol{AV}_{\text{lin}}[\beta_j; \eta^2].$$

The subscript "lin" refers to validity of this asymptotic variance under the assumption-loaded model-trusting framework of linear models theory.

11.4 *RAV*—Ratio of Proper and Improper Asymptotic Variances

To examine the discrepancies between proper and improper asymptotic variances, we form their ratio, which results in the following elegant functional of the conditional MSE and the squared adjusted regressor:

DEFINITION. Ratio of Asymptotic Variances.

$$\boxed{\begin{aligned}\boldsymbol{RAV}[\beta_j, m^2] &\triangleq \frac{\boldsymbol{AV}_{\text{lean}}[\beta_j, m^2]}{\boldsymbol{AV}_{\text{lin}}[\beta_j, m^2]} \\ &= \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})X_{j\bullet}^2]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]\boldsymbol{E}[X_{j\bullet}^2]}.\end{aligned}}$$

In order to examine the effect of heteroskedasticities and nonlinearities on the discrepancies separately, one can also define $\boldsymbol{RAV}[\beta_j, \sigma^2]$ and $\boldsymbol{RAV}[\beta_j, \eta^2]$. By the decomposition lemma in Appendix E.2, $\boldsymbol{RAV}[\beta_j, m^2]$ is a weighted mixture of these two terms. Belaboring the obvious, the interpretation of the $\boldsymbol{RAV}$ is this: If $\boldsymbol{RAV}[\beta_j, m^2] > 1$, then $\hat{\boldsymbol{SE}}_{\text{lin}}[\hat{\beta}_j]$ is asymptotically too small/optimistic; if $\boldsymbol{RAV}[\beta_j, m^2] = 1$, it is asymptotically correct; else it is asymptotically too large/pessimistic.

We will later have use for the following sufficient conditions for $\boldsymbol{RAV} = 1$. The second condition says that when the population residual δ is a traditional error term (but not necessarily Gaussian), the usual standard error of linear models theory is asymptotically correct.

LEMMA 11.1. *Sufficient conditions for* $\boldsymbol{RAV}[\beta_j, m^2] = 1$ *are the following*:

(a) $m^2(\vec{\boldsymbol{X}}) = m_0^2$ *is constant*.
(b) δ^2 *and* $X_{j\bullet}{}^2$ *are independent*.

PROOF. Assertion (a) is immediate from the definition of $\boldsymbol{RAV}[\beta_j, m^2]$. Assertion (b): The numerator of $\boldsymbol{RAV}[\beta_j, m^2]$ is $\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2] = \boldsymbol{E}[\delta^2 X_{j\bullet}{}^2] = \boldsymbol{E}[\delta^2]\boldsymbol{E}[X_{j\bullet}{}^2]$, hence equals the denominator. □

The ratio $\boldsymbol{RAV}[\beta_j, m^2]$ is the inner product between the random variables

$$\frac{m^2(\vec{\boldsymbol{X}})}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]} \quad \text{and} \quad \frac{X_{j\bullet}^2}{\boldsymbol{E}[X_{j\bullet}^2]}.$$

It is *not* a correlation as both $m^2(\vec{\boldsymbol{X}})$ and $X_{j\bullet}^2$ are L_1-normalized; a noncentered correlation would require L_2-normalization with denominators $\boldsymbol{E}[m^4(\vec{\boldsymbol{X}})]^{1/2}$ and $\boldsymbol{E}[X_{j\bullet}^4]^{1/2}$, respectively. Its upper bound is obviously not $+1$ but rather ∞.

11.5 The Range of *RAV*

The analysis of the ***RAV*** is simplified by conditioning $m^2(\vec{X})$ on $X_{j\bullet}^2$.

DEFINITION AND LEMMA 11.1. Letting

$$m_j^2(X_{j\bullet}^2) \triangleq \boldsymbol{E}[m^2(\vec{X})|X_{j\bullet}^2],$$

we have

$$\boldsymbol{RAV}[\beta_j, m^2] = \boldsymbol{RAV}[\beta_j, m_j^2].$$

Thus the analysis of the ***RAV*** is reduced to single squared adjusted regressors $X_{j\bullet}^2$. This fact lends itself to simple case studies and graphical illustrations.

Next, we describe the extremes of the ***RAV*** over scenarios of $m^2(\vec{X})$ or, by Lemma 11.1, of $m_j^2(X_{j\bullet}^2)$.

PROPOSITION. *If* $\boldsymbol{E}[X_{j\bullet}^2] < \infty$ *and* $X_{j\bullet}^2$ *has unbounded support, then*

$$\sup_{m_j^2} \boldsymbol{RAV}[\beta_j, m_j^2] = \infty.$$

If $\boldsymbol{E}[X_{j\bullet}^2] < \infty$ *and* $X_{j\bullet}^2$ *has* 0 *in its support, then*

$$\inf_{m_j^2} \boldsymbol{RAV}[\beta_j, m_j^2] = 0.$$

Thus, when the adjusted regressor distribution is unbounded, the usual standard error can be too small to any degree. Conversely, if the adjusted regressor is not bounded away from zero, it can be too large to any degree.

What shapes of $m_j^2(X_{j\bullet}^2)$ approximate these extremes? The answer can be gleaned from Figure 6 which illustrates the proposition for normally distributed $X_{j\bullet}$: If *nonlinearities and/or heteroskedasticities blow up* ...

- in the *tails* of the $X_{j\bullet}$ distribution, then ***RAV*** takes on *large* values;
- in the *center* of the $X_{j\bullet}$ distribution, then ***RAV*** takes on *small* values.

The proof in Appendix E.3 bears this out. As the main concern is with usual standard errors that are too small, $\boldsymbol{RAV} > 1$, the proposition indicates that $X_{j\bullet}$-distributions with bounded support enjoy some protection from the worst case.

11.6 Illustration of Factors That Drive the *RAV*

We further analyze the ***RAV*** in terms of the constituents of $m_j^2(X_{j\bullet}^2)$, conditional variance and squared nonlinearity, as functions of $X_{j\bullet}^2$:

$$\begin{aligned} \sigma_j^2(X_{j\bullet}^2) &= \boldsymbol{E}[\sigma^2(\vec{X})|X_{j\bullet}^2] \quad \text{and} \\ \eta_j^2(X_{j\bullet}^2) &= \boldsymbol{E}[\eta^2(\vec{X})|X_{j\bullet}^2]. \end{aligned} \tag{23}$$

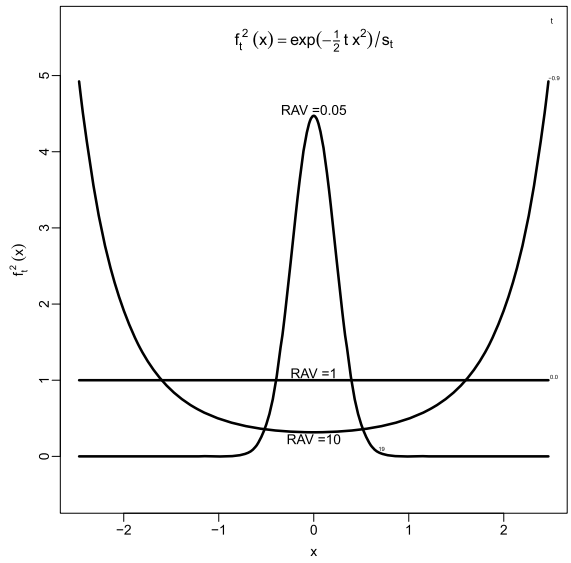

FIG. 6. *A family of functions* $f_t^2(x)$ *that can be interpreted as conditional MSEs* $m_j^2(X_{j\bullet}^2)$, *heteroskedasticities* $\sigma_j^2(X_{j\bullet}^2)$ *or squared nonlinearities* $\eta_j^2(X_{j\bullet}^2)$ (*shown as functions of* $x = X_{j\bullet}$ *rather than* $X_{j\bullet}^2$): *The family interpolates* ***RAV*** *from* 0 *to* ∞ *for* $x = X_{j\bullet} \sim N(0,1)$. *The three solid black curves show* $f_t^2(x)$ *that result in RAV* $= 0.05$, 1, *and* 10. (*See Appendix E.4 for details.*) $\boldsymbol{RAV} = \infty$ *is approached as* $f_t^2(x)$ *bends ever more strongly in the tails of the x-distribution.* $\boldsymbol{RAV} = 0$ *is approached by an ever stronger spike in the center of the x-distribution.*

For qualitative insights into the drivers of the ***RAV***, we translate (23) to concrete data scenarios. Figure 7 shows three noise scenarios and Figure 8 three nonlinearity scenarios. The illustrated effects will both be present to degrees in real data. Their combined effect is described by a decomposition lemma in Appendix E.2: $\boldsymbol{RAV}[\beta_j, m_j^2]$ is a weighted mixture of $\boldsymbol{RAV}[\beta_j, \sigma_j^2]$ and $\boldsymbol{RAV}[\beta_j, \eta_j^2]$. Therefore:

- Heteroskedasticities with large $\sigma_j^2(X_{j\bullet}^2)$ in the tail of $X_{j\bullet}{}^2$ produce an upward contribution to $\boldsymbol{RAV}[\beta_j, m_j^2]$; heteroskedasticities with large $\sigma_j^2(X_{j\bullet}^2)$ near $X_{j\bullet}^2 = 0$ imply a downward contribution to $\boldsymbol{RAV}[\beta_j, m_j^2]$.
- Nonlinearities with large average values $\eta_j^2(X_{j\bullet}^2)$ in the tail of $X_{j\bullet}^2$ imply an upward contribution to $\boldsymbol{RAV}[\beta_j, m_j^2]$; nonlinearities with large $\eta_j^2(X_{j\bullet}^2)$ concentrated near $X_{j\bullet}^2 = 0$ imply a downward contribution to $\boldsymbol{RAV}[\beta_j, m_j^2]$.

These facts also suggest that large values $\boldsymbol{RAV} > 1$ should occur more often than small values $\boldsymbol{RAV} < 1$

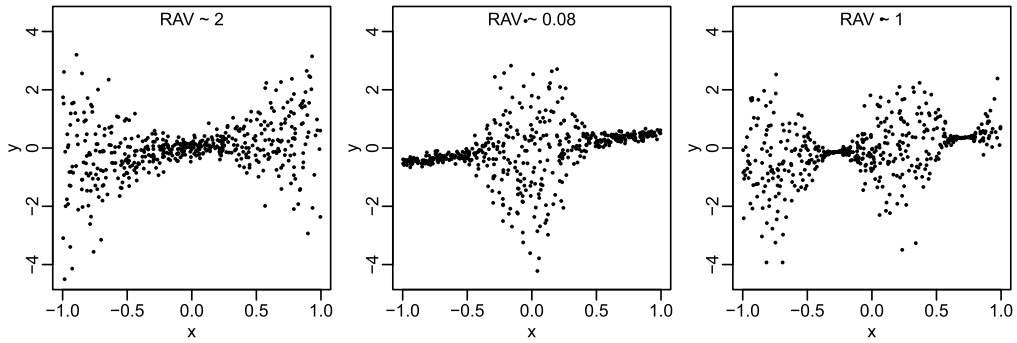

FIG. 7. *The effect of heteroskedasticity on the sampling variability of slope estimates*: *How does the treatment of the heteroskedasticities as homoskedastic affect statistical inference? Left*: *High noise variance in the tails of the regressor distribution elevates the true sampling variability of the slope estimate above the usual standard error*: $\boldsymbol{RAV}[\beta_j, \sigma^2] > 1$. *Center*: *High noise variance near the center of the regressor distribution lowers the true sampling variability of the slope estimate below the usual standard error*: $\boldsymbol{RAV}[\beta_j, \sigma^2] < 1$. *Right*: *The noise variance oscillates in such a way that the usual standard error is coincidentally correct* ($\boldsymbol{RAV}[\beta_j, \sigma^2] = 1$).

because large conditional variances as well as nonlinearities are often more pronounced in the extremes of regressor distributions, not their centers. This is most natural for nonlinearities which are often convex or concave. Also, it follows from the ***RAV*** decomposition lemma (Appendix E.2) that either of $\boldsymbol{RAV}[\beta_j, \sigma_j^2]$ or $\boldsymbol{RAV}[\beta_j, \eta_j^2]$ is able to single-handedly pull $\boldsymbol{RAV}[\beta_j, m_j^2]$ to $+\infty$, whereas both have to be close to zero to pull $\boldsymbol{RAV}[\beta_j, m_j^2]$ toward zero. These heuristics support the observation that in practice $\hat{\boldsymbol{SE}}_{\text{lin}}$ is more often too small than too large compared to the asymptotically correct $\hat{\boldsymbol{SE}}_{\text{sand}}$.

12. SANDWICH ESTIMATORS IN ADJUSTED FORM AND A *RAV* TEST

The goal here is to write the ***RAV*** in adjustment form and estimate it with plug-in for use as a test statistic to decide whether the usual standard error is adequate. We will obtain one test per regressor. The proposal is related to the class of "misspecification tests" for which there exists a literature starting with Hausman (1978) and continuing with White (1980a, 1980b, 1981, 1982) and others. These tests are largely global rather than coefficient-specific, which ours is. The test proposed here has similarities to White's (1982, Section 4) "information matrix test" which compares two types of

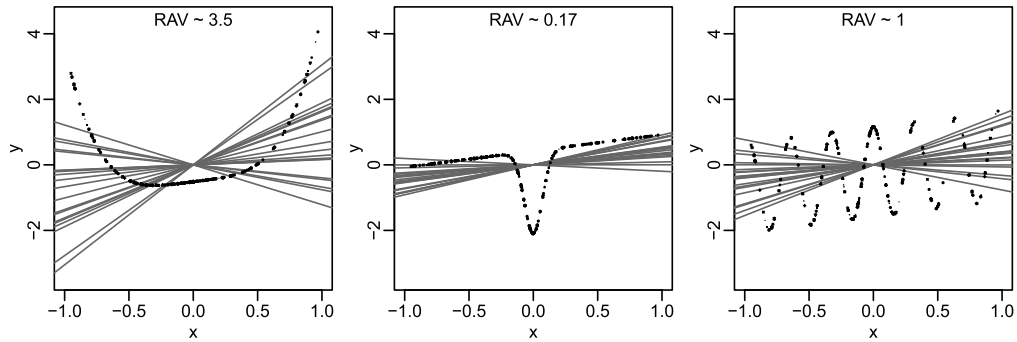

FIG. 8. *The effect of nonlinearities on the sampling variability of slope estimates*: *The three plots show three different noise-free nonlinearities*; *each plot shows for one nonlinearity* 20 *overplotted datasets of size* $N = 10$ *and their fitted lines through the origin. The question is how the misinterpretation of the nonlinearities as homoskedastic random errors affects statistical inference. Left*: *Strong nonlinearity in the tails of the regressor distribution elevates the true sampling variability of the slope estimate above the usual standard error* ($\boldsymbol{RAV}[\beta_j, \eta^2] > 1$). *Center*: *Strong nonlinearity near the center of the regressor distribution lowers the true sampling variability of the slope estimate below the usual standard error* ($\boldsymbol{RAV}[\beta_j, \eta^2] < 1$). *Right*: *An oscillating nonlinearity mimics homoskedastic random error to make the usual standard error coincidentally correct* ($\boldsymbol{RAV}[\beta_j, \eta^2] = 1$). *Caveat*: *These are cartoons illustrating potential causes of standard error discrepancies. Nonlinearities may not be detectable in actual data in the presence of noise and other regressors.*

information matrices globally, while we compare two types of standard errors, one coefficient at a time.

12.1 Sandwich Estimators in Adjustment Form and the $\hat{\boldsymbol{RAV}}_j$ Test Statistic

The adjustment versions of the asymptotic variances in the CLTs of Corollary 11.1 can be used to rewrite the sandwich estimator by replacing expectations $\boldsymbol{E}[\ldots]$ with means $\hat{\boldsymbol{E}}[\ldots]$, $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$, $X_{j\bullet}$ with $\mathbf{X}_{j\hat{\bullet}}$, and rescaling by N:

$$\begin{aligned}\hat{\boldsymbol{SE}}_{\text{sand}}[\hat{\beta}_j]^2 &= \frac{1}{N}\frac{\hat{\boldsymbol{E}}[(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2]}{\hat{\boldsymbol{E}}[X_{j\hat{\bullet}}{}^2]^2} \\ &= \frac{\langle \boldsymbol{r}^2, \mathbf{X}_{j\hat{\bullet}}{}^2\rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^4}.\end{aligned} \tag{24}$$

The squaring of N-vectors is meant to be coordinatewise. Formula (24) is algebraically equivalent to the diagonal elements of (18).

To match the raw plug-in form of the sandwich estimator (24), we use the plug-in version of the standard error estimator of linear models theory, the only difference being division by N rather than $N-p-1$:

$$\hat{\boldsymbol{SE}}_{\text{lin}}[\hat{\beta}_j]^2 = \frac{1}{N}\frac{\hat{\boldsymbol{E}}[(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2]}{\hat{\boldsymbol{E}}[X_{j\hat{\bullet}}{}^2]} = \frac{1}{N}\frac{\|\boldsymbol{r}\|^2}{\|\mathbf{X}_{j\hat{\bullet}}\|^2}. \tag{25}$$

Thus the plug-in estimate of $\boldsymbol{RAV}[\beta_j, m^2]$ is

$$\begin{aligned}\hat{\boldsymbol{RAV}}_j &\triangleq \frac{\hat{\boldsymbol{E}}[(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2]}{\hat{\boldsymbol{E}}[(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2]\hat{\boldsymbol{E}}[X_{j\hat{\bullet}}{}^2]} \\ &= N\frac{\langle \boldsymbol{r}^2, \mathbf{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{r}\|^2\|\mathbf{X}_{j\hat{\bullet}}\|^2}.\end{aligned} \tag{26}$$

This is the proposed test statistic. Analogous to the population-level $\boldsymbol{RAV}[\beta_j, m^2]$, the sample-level $\hat{\boldsymbol{RAV}}_j$ responds to associations between squared residuals and squared adjusted regressors.

12.2 The Asymptotic Null Distribution of the *RAV* Test Statistic

Here is an asymptotic result that would be expected to yield approximate inference under a null hypothesis that implies $\boldsymbol{RAV}[\beta_j, m^2] = 1$ based on Lemma 11.1.

PROPOSITION. *Under the null hypothesis H_0 that the population residuals δ and the adjusted regressor $X_{j\bullet}$ are independent, it holds*:

$$\begin{aligned}&N^{1/2}(\hat{\boldsymbol{RAV}}_j - 1) \\ &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\boldsymbol{E}[\delta^4]}{\boldsymbol{E}[\delta^2]^2}\frac{\boldsymbol{E}[X_{j\bullet}{}^4]}{\boldsymbol{E}[X_{j\bullet}^2]^2} - 1\right).\end{aligned} \tag{27}$$

As always, we ignore technical assumptions. A proof outline is in Appendix E.5.

The asymptotic variance of $\hat{\boldsymbol{RAV}}_j$ under H_0 is driven by the standardized fourth moments or the kurtoses (= same -3) of δ and $X_{j\bullet}$. Some observations:

1. The larger the kurtosis of population residuals δ and/or adjusted regressors $X_{j\bullet}$, the less likely is detection of first- and second-order model misspecification resulting in standard error discrepancies.
2. As standardized fourth moments are always ≥ 1 by Jensen's inequality, the asymptotic variance is ≥ 0, as it should be. The asymptotic variance vanishes iff the minimal standardized fourth moment is $+1$ for both δ and $X_{j\bullet}$, hence both have symmetric two-point distributions (as both are centered). For such $X_{j\bullet}$, it holds $\boldsymbol{RAV}[\beta_j, m^2] = 1$ by Proposition E.3 in the Appendix.
3. A test of the stronger H_0 that includes normality of δ is obtained by setting $\boldsymbol{E}[\delta^4]/\boldsymbol{E}[\delta^2]^2 = 3$ rather than estimating it. The result, however, is an overly sensitive nonnormality test much of the time, which does not seem useful as nonnormality can be diagnosed and tested by other means.

12.3 An Approximate Permutation Distribution for the *RAV* Test Statistic

The asymptotic result of the proposition in Section 12.2 provides qualitative insights, but it is not suitable for practical application because the null distribution of $\hat{\boldsymbol{RAV}}_j$ can be very nonnormal for finite N, and this in ways that are not easily overcome with simple tools such as nonlinear transformations. Another approach to null distributions for finite N is needed, and it is available in the form of an approximate permutation test because H_0 is just a null hypothesis of independence, here between δ and $X_{j\bullet}$. The test is not exact, requiring $N \gg p$, because population residuals δ_i must be estimated with sample residuals r_i and population adjusted regressor values $X_{i,j\bullet}$ with sample adjusted analogs $X_{i,j\hat{\bullet}}$. The permutation simulation is cheap: Once coordinatewise squared vectors $\boldsymbol{r}^2$ and $\mathbf{X}_{j\hat{\bullet}}{}^2$ are formed, a draw from the conditional null distribution of $\hat{\boldsymbol{RAV}}_j$ is obtained by randomly permuting one of the vectors and forming the inner product with the other, rescaled by a permutation-invariant factor $N/(\|\boldsymbol{r}\|^2\|\mathbf{X}_{j\hat{\bullet}}\|^2)$. A retention interval should be formed directly from the $\alpha/2$ and $1-\alpha/2$ quantiles of the permutation distribution to account for distributional asymmetries. The permutation distribution also

TABLE 4

LA Homeless data: Permutation inference for $\hat{\boldsymbol{RAV}}_j$ (10,000 *permutations*). *The value of* $\hat{\boldsymbol{RAV}}_j$ *for* `PercIndustrial` *detects a statistically significant difference between* $\boldsymbol{SE}_{\text{lin}}$ *and* $\boldsymbol{SE}_{\text{sand}}$ *for this regressor. See also Table* 4 *in the Appendix for the Boston Housing data where significant differences are detected for six of* 13 *regressors*

	$\hat{\beta}_j$	$\boldsymbol{SE}_{\text{lin}}$	$\boldsymbol{SE}_{\text{sand}}$	$\hat{\boldsymbol{RAV}}_j$	2.5% Perm.	97.5% Perm.
`(Intercept)`	0.760	22.767	16.209	0.495	0.354	3.182
`MedianIncome ($K)`	−0.183	0.187	0.108	0.318	0.274	5.059
`PercVacant`	4.629	0.901	1.363	2.071	0.303	3.823
`PercMinority`	0.123	0.176	0.164	0.860	0.403	2.238
`PercResidential`	−0.050	0.171	0.111	0.406	0.369	3.058
`PercCommercial`	0.737	0.273	0.397	2.046	0.355	3.073
`PercIndustrial`	0.905	0.321	0.592	3.289*	0.323	3.215

yields an easy diagnostic of nonnormality (see Appendix F for examples). Finally, by applying permutation simulations simultaneously to ***RAV*** statistics of multiple regressors, one can calibrate the retention intervals to control familywise error. See Table 4 (as well as Table 4 in Appendix B) for examples of ***RAV*** tests.

13. ISSUES WITH MODEL-ROBUST STANDARD ERRORS

Model-robustness is a highly desirable property, but as always there is no free lunch. Kauermann and Carroll (2001) have shown that a cost of the sandwich estimator can be ***inefficiency when the assumed model is correct***. Sandwich estimators should be accurate only when the sample size is sufficiently large.

Another cost associated with the sandwich estimator is ***nonrobustness in the sense of robust statistics*** (Huber and Ronchetti, 2009; Hampel et al. 1986), meaning strong sensitivity to heavy-tailed distributions: The statistic $\hat{\boldsymbol{SE}}^2_{\text{sand}}[\hat{\beta}_j]$ of (24) is a ratio of fourth-order quantities of the data, whereas $\hat{\boldsymbol{SE}}^2_{\text{lin}}[\hat{\beta}_j]$ of (25) is "only" a ratio of second order quantities.[13] The two types of robustness are in conflict: Model-robust standard error estimators are highly nonrobust to heavy tails compared to their model-trusting analogs. This is a large issue which we can only raise but not solve. Here are some observations and suggestions:

- Classical robust regression may confer partial robustness to the sandwich standard error as it caps residuals with a bounded ψ function, thereby addressing robustness to heavy tails in the vertical (y) direction. Anecdotal evidence suggests partial benefits. In the LA Homeless data, for example, when comparing bootstrap standard errors and standard errors reported by the **R** (2008) software (function `lmrob` in package `robustbase`), we observed ratios $\frac{\boldsymbol{SE}_{\text{boot}}}{\boldsymbol{SE}_{\text{lmrob}}}$ of 1.470 and 0.957 for the coefficients of `PercVacant` and `PercIndustrial`, respectively. For linear OLS, the corresponding ratios $\frac{\boldsymbol{SE}_{\text{boot}}}{\boldsymbol{SE}_{\text{lin}}}$ in Table 1 were 1.513 and 1.843, respectively. Thus the roughly 50% discrepancy for `PercVacant` persists, but the 80% discrepancy for `PercIndustrial` is completely corrected.
- Heavy-tail robustness in the horizontal ($\vec{\boldsymbol{x}}$) direction can be achieved with bounded-influence regression (e.g., Krasker and Welsch, 1982, and references therein) which downweights observations in high-leverage positions.
- Robustness to horizontally heavy tails can also be addressed by transforming the regressor variables to bounded ranges (though this changes the meaning of the slopes). Taking a cue from Proposition E.3 in the Appendix, one might search for transformations that obviate the need for a model-robust standard error in the first place.

To illustrate the last point, we transformed the regressors of the LA Homeless data with their empirical cdfs to achieve approximately uniform marginal distributions. The transformed data are no longer i.i.d., but the point is to examine the effect of transforming the regressors to a finite range. As a result, shown in Table 1 of Appendix A, the discrepancies between sandwich and usual standard errors have all but disappeared. The same drastic effect is not seen in the Boston Housing data (Appendix B, Table 3), although the discrepancies are greatly reduced, too.

[13] Note we are here concerned with nonrobustness of standard error estimates, not parameter estimates.

14. SUMMARY AND OUTLOOK

We explored for linear OLS the idea that statistical models imply "simplification and idealization" (Cox, 1995), and hence should be treated as approximations rather than truths. The implications are many: (1) Slope parameters need to be reinterpreted as statistical functionals $\boldsymbol{\beta}(\boldsymbol{P}_{Y,\vec{X}})$ arising from best-approximating linear equations to essentially arbitrary conditional mean functions $\mu(\vec{\boldsymbol{X}})$; (2) the presence of nonlinearity $\eta(\vec{\boldsymbol{X}})$ requires new interpretations of slope parameters and their estimates; (3) regressors are no longer ancillary for the slope parameters; hence (4) conditioning on the regressors is not justified and regressors must be treated as random, arising from a regressor distribution $\boldsymbol{P}_{\vec{X}}$; (5) nonlinearity causes slope parameters to depend not only on the conditional response distribution $\boldsymbol{P}_{Y|\vec{X}}$ but on the regressor distribution $\boldsymbol{P}_{\vec{X}}$ as well; (6) nonlinearity causes randomness in the regressors $\vec{\boldsymbol{X}}$ to generate sampling variation in slope estimates of order $N^{-1/2}$; (7) sampling variability due to $Y|\vec{\boldsymbol{X}}$ and due to $\vec{\boldsymbol{X}}$ are asymptotically correctly captured by model-robust standard error estimates from the $x{-}y$ bootstrap and sandwich plug-in, the latter being a limiting case of the former; (8) the factors that render the usual standard error of a slope too liberal are strong nonlinearity and/or large noise variance in the extremes of the adjusted regressor; (9) validity of the usual standard error varies from slope to slope but can be tested with a slope-specific test; (10) unresolved remains the problem that model-robustness and classical heavy-tail robustness of standard error estimates appear to be in conflict with each other.

A vexing item in this list is (2): What is the meaning of a slope in the presence of nonlinearity? We gave an answer in terms of average observed slopes, but this issue may remain controversial. Yet, the traditional interpretation of slopes should be even more controversial because the notion of "average difference in the response for a unit difference in the regressor, ceteris paribus," tacitly assumes the fitted linear equation to be correctly specified. It remains correct if "in the response" is replaced by "in the best linear approximation," but this correction may leave some dissatisfied as well. Yet, misspecification is often a fact, as when simple models are needed for substantive reasons or for communication with consumers of statistical analysis. It may then be prudent to use interpretations and inferences that do not assume correct specification.

Since White's seminal work, research into misspecification has progressed far in addressing specific classes of misspecification: dependencies, heteroskedasticities and nonlinearities. A generalization of White's sandwich estimator to time series dependence in regression is the "heteroskedasticity and autocorrelation consistent" (HAC) estimator of standard error by Newey and West (1987). Structured second order misspecification such as over/underdispersion have been addressed with quasilikelihood. Intra-cluster dependencies in clustered (e.g., longitudinal) data have been addressed with generalized estimating equations (GEE) where the sandwich estimator is in common use, as it is in the generalized method of moments (GMM) literature. Finally, nonlinearities have been modeled with specific function classes or estimated nonparametrically with, for example, additive models, spline and kernel methods, and tree-based fitting. In spite of these advances, in finite data not all possibilities of misspecification can be approached simultaneously, and there still arises a need for model-robust inference.

There exist, finally, areas that frequently rely on model-trusting theory:

- Bayes inference based on uninformative priors is asymptotically equivalent to model-trusting frequentist inference (Hartigan, 1983, Ch. 11). It should be reasonable to ask how much inferences from Bayesian models are adversely affected by misspecification. After the early work by Berk (1966, 1970) we find some more recent promising developments: Szpiro, Rice and Lumley (2010) derive a sandwich estimator from Bayesian assumptions, and a lively discussion of misspecification from a Bayesian perspective involved Walker (2013), De Blasi (2013), Hoff and Wakefield (2013) and O'Hagan (2013), who provide further references.
- High-dimensional inference is the subject of a large literature that often relies on the assumptions of linearity, homoskedasticity as well as normality. It may be uncertain whether procedures proposed in this area are model-robust. Recently, however, attention to the issue started to be paid by Bühlmann and van de Geer (2015). Relevant is also the incorporation of ideas from classical robust statistics by, for example, El Karoui et al. (2013), Donoho and Montanari (2014) and Loh (2017).

In summary, while interesting developments are in progress, there remain open problems, especially in some of today's most lively research areas. Even in the non-Bayesian and low-dimensional domain there remains the conflict between model-robustness and classical robustness. The implications of statistical models viewed as approximations are not yet satisfactorily realized.

ACKNOWLEDGMENTS

A. Buja, L. Brown, E. Pitkin, L. Zhao and K. Zhang were supported in part by NSF Grants DMS-10-07657 and DMS-1310795. E. George was supported in part by NSF Grant DMS-14-06563.

SUPPLEMENTARY MATERIAL

Supplement to "Models as Approximations I: Consequences Illustrated with Linear Regression" (DOI: 10.1214/18-STS693SUPP; .pdf). This supplement contains Appendices A-F.

REFERENCES

BERK, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Stat.* **37** 51–58. MR0189176

BERK, R. H. (1970). Consistency a posteriori. *Ann. Math. Stat.* **41** 894–906. MR0266356

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

BERK, R., KRIEGLER, B. and YLVISAKER, D. (2008). Counting the homeless in Los Angeles county. In *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Stat. (IMS) Collect.* **2** 127–141. IMS, Beachwood, OH. MR2459951

BERMAN, M. (1988). A theorem of Jacobi and its generalization. *Biometrika* **75** 779–783. MR0995120

BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statist. Sinica* **7** 1–31. MR1441142

BOX, G. E. P.(1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics: Proceedings of a Workshop* (R. L. Launer and G. N. Wilkinson, eds.) Academic Press (Elsevier), Amsterdam.

BÜHLMANN, P. and VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Stat.* **9** 1449–1473. MR3367666

BUJA, A., BROWN, L., BERK, R., GEORGE, E., PITKIN, E., TRASKIN, M., ZHANG, K., and ZHAO, L. (2019). Supplement to "Models as Approximations I: Consequences Illustrated with Linear Regression." 10.1214/18-STS693SUPP.

COX, D. R. (1962). Further results on tests of separate families of hypotheses (1962). *J. Roy. Statist. Soc. Ser. B* **24** 406–424.

COX, D. R. (1995). Discussion of Chatfield (1995). *J. Roy. Statist. Soc. Ser. A* **158** 455–456.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. CRC Press, London. MR0370837

DAVIES, L. (2014). *Data Analysis and Approximate Models. Monographs on Statistics and Applied Probability* **133**. CRC Press, Boca Raton, FL. MR3241514

DE BLASI, P. (2013). Discussion on article "Bayesian inference with misspecified models" by Stephen G. Walker. *J. Statist. Plann. Inference* **143** 1634–1637. MR3082221

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

DONOHO, D. and MONTANARI, A. (2014). Variance Breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. Available at arXiv:1503.02106.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. SIAM, Philadelphia, PA. MR0659849

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903

EICKER, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Stat.* **34** 447–456. MR0148177

EL KAROUI, N., BEAN, D., BICKEL, P. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci.* **110** 14563–14568.

FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228. MR0630104

FREEDMAN, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *Amer. Statist.* **60** 299–302. MR2291297

GELMAN, A. and PARK, D. K. (2009). Splitting a predictor at the upper quarter or third and the lower quarter or third. *Amer. Statist.* **63** 1–8. MR2655696

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion. Springer Series in Statistics*. Springer, New York. MR1145237

HALL, A. R. (2005). *Generalized Method of Moments. Advanced Texts in Econometrics*. Oxford Univ. Press, Oxford. MR2135106

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, New York. MR0829458

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

HARTIGAN, J. A. (1983). Asymptotic normality of posterior distributions. In *Bayes Theory. Springer Series in Statistics* 107–118. Springer, New York, NY.

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271. MR0513692

HINKLEY, D. V. (1977). Jacknifing in unbalanced situations. *Technometrics* **19** 285–292. MR0458734

HOFF, P. and WAKEFIELD, J. (2013). Bayesian sandwich posteriors for pseudo-true parameters. A discussion of "Bayesian inference with misspecified models" by Stephen Walker. *J. Statist. Plann. Inference* **143** 1638–1642. MR3082222

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

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795

KAUERMANN, G. and CARROLL, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.* **96** 1387–1396. MR1946584

KRASKER, W. S. and WELSCH, R. E. (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.* **77** 595–604. MR0675886

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948

LEHMANN, E. L. and ROMANO, J. P. (2008). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR1481711

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

LOH, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403

LONG, J. S. and ERVIN, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear model. *Amer. Statist.* **54** 217–224.

MACKINNON, J. and WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econometrics* **29** 305–325.

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176

MAMMEN, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Statist.* **24** 307–335. MR1389892

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

NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864

O'HAGAN, A. (2013). Bayesian inference with misspecified models: Inference about what?. *J. Statist. Plann. Inference* **143** 1643–1648. MR3082223

OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton, FL.

POLITIS, D. N. and ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050. MR1329181

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.

STIGLER, S. M. (2001). Ancillary history. In *State of the Art in Probability and Statistics*: *Festschrift for W. R. van Zwet* (M. DeGunst, C. Klaassen and A. van der Vaart, eds.) 555–567.

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

WALKER, S. G. (2013). Bayesian inference with misspecified models. *J. Statist. Plann. Inference* **143** 1621–1633. MR3082220

WASSERMAN, L. (2011). Low assumptions, high dimensions. *Ration. Mark. Moral.* **2** 201–209.

WEBER, N. C. (1986). The jackknife and heteroskedasticity: Consistent variance estimation for regression models. *Econom. Lett.* **20** 161–163. MR0830028

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

WHITE, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc.* **76** 419–433. MR0624344

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

WHITE, H. (1994). *Estimation, Inference and Specification Analysis*. *Econometric Society Monographs* **22**. Cambridge Univ. Press, Cambridge. MR1292251

WU, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1350. MR0868303