Submitted to Statistical Science

Models as Approximations, Part I: A Conspiracy of Nonlinearity and Random Regressors in Linear Regression

Andreas Buja*,†,‡**, Richard Berk**‡**, Lawrence Brown***,‡**, Edward George**‡**, Emil Pitkin***,‡**, Mikhail Traskin**§ **, Linda Zhao***,‡ **and Kai Zhang***,¶

Wharton – University of Pennsylvania‡ and Citadel§ and UNC at Chapel Hill¶

Dedicated to Halbert White (†2012)

Abstract.

More than thirty years ago 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 regressors are no longer ancillary, hence can't be treated as fixed. The consequences are severe: (1) the regressor distribution affects the slope parameters, and (2) randomness of the regressors conspires with the nonlinearity to create sampling variability in slope estimates — even in the complete absence of error. For these observations to make sense it is necessary to re-interpret population slopes and view them not as parameters in a generative model but as statistical functionals associated with OLS fitting as it applies to largely arbitrary joint x-y distributions. In such a "model-robust" approach to linear regression, the meaning of slope parameters needs to be rethought and inference needs to be based on model-robust standard errors that can be estimated with sandwich plug-in estimators or with the x-y bootstrap. Theoretically, model-robust and model-trusting standard errors can deviate by arbitrary magnitudes either way. In practice, a diagnostic test can be used to detect significant deviations on a per-slope basis.

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

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

Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 (e-mail: buja.at.wharton@gmail.com). – Citadel, Chicago. – Dept. of Statistics & Operations Research, 306 Hanes Hall, CB#3260, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260.

*Supported in part by NSF Grant DMS-10-07657.

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

1. INTRODUCTION

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

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

The right hand side 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:

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

where $\hat{\boldsymbol{D}}$ is diagonal with $\hat{\boldsymbol{D}}_{ii} = r_i^2$. Standard error estimates are obtained by $\hat{\boldsymbol{SE}}_{sand}[\,\hat{\beta}_j\,] = \hat{\boldsymbol{V}}_{sand}[\,\hat{\boldsymbol{\beta}}\,]_{jj}^{1/2}$. 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. 162-3) less widely read articles: the sandwich estimator of standard error is asymptotically correct even in the presence of nonlinearity:

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

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 random.* The sandwich estimate 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 all of its own, even in the absence of noise.

For a quick and intuitive grasp of these "conspiracy effects", the reader may peruse Figure 2 below for effect (1) and Figure 4 for effect (2). A more striking illustration of effect (2) 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")
```

The code generates an animation that shows how variability in slope and intercept estimates arises under repeated sampling of datasets when the regressors are random and the true response is a nonlinear function of the regressors, even in the complete absence of error.

Side remarks:

- The term "nonlinearity" is meant in the sense of (3), 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. We distinguish between "regressors" and "independent variables": Multiple regressors may be functions of one or more independent variables.

- The sandwich estimator (2) 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).

From the sandwich estimator (2), the usual *model-trusting* estimator is obtained by collapsing the sandwich form assuming homoskedasticity:

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] := (\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2, \qquad \hat{\sigma}^2 = \|\boldsymbol{r}\|^2/(N{-}p{-}1).$$

This yields finite-sample unbiased squared standard error estimators $\hat{\boldsymbol{SE}}^2_{lin}[\,\hat{\beta}_j\,] = \hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,]_{jj}$ 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 iid 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 and related questions is a first goal of this article. An established answer is that parameters can be re-interpreted as *statistical functionals* $\boldsymbol{\beta}(\boldsymbol{P})$ defined on a large nonparametric class of joint distributions $\boldsymbol{P} = \boldsymbol{P}(dy, d\vec{\boldsymbol{x}})$ through best approximation (Section 3). 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 *case-wise* and *pairwise slopes* (Section 10).

A second goal of this article is to discuss the role of the regressors when they are random. Based on an ancillarity argument, *model-trusting* theories tend to condition on the regressors and treat them as fixed (Cox and Hinkley 1974, p. 32f, Lehmann and Romano 2008, p. 395ff). It will be shown that in a *model-robust* theory the ancillarity principle is violated in the sense that population parameters depend on the distribution of the regressors (Section 4). The consequences of this fact are vast and will be elaborated in Part II for largely arbitrary types of regression that permit an interpretation of parameters as statistical functionals.

A third goal of this article is to connect the sandwich estimator and the "x-y bootstrap" which resamples observations $(\vec{\boldsymbol{x}}_i', y_i)$. The better known "residual bootstrap" resamples residuals r_i. Theory exists for both (Freedman (1981) and Mammen (1993), for example), but only the x-y bootstrap is model-robust and solves the same problem as the sandwich estimator. Indeed, it will be shown that the sandwich estimator is a limiting case of the x-y bootstrap (Section 8).

A fourth goal of this article is to practically (Section 2) and theoretically (Section 11) compare model-robust and model-trusting estimators in the case of linear OLS. We define a ratio of asymptotic variances — "***RAV***" for short — that describes the discrepancies between the two standard errors in the asymptotic limit. If ***RAV***$\neq$1, it is model-robust estimators (sandwich or x-y bootstrap) that are asymptotically correct, and the usual model-trusting standard error is indeed asymptotically incorrect. The ***RAV*** can range from 0 to ∞ under scenarios that illustrate how model deviations can invalidate the usual standard error.

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). It should be noted that there are misspecifications that do not invalidate the usual model-trusting standard error.

A final goal is to briefly discuss issues with the sandwich estimator (Section 13): When models are well-specified, it can be inefficient. We additionally point out that it is also very non-robust in the sense of sensitivity to outlying observations. This is an issue we consider an open problem. [To make sense of this brief forward pointer, the following distinctions are needed: (1) classical robustness to outlying observations 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 that are the subject of the present article.]

Throughout we use precise notation for clarity, yet this article is not very technical. The majority of results is 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 Part I of this two-part series is limited to linear models or, more precisely, OLS as the criterion that defines best fits, both for populations and for data. OLS allows the most explicit calculations and lucid interpretations of the issues. Part II will be concerned with an analysis of the notion of mis- and well-specification for largely arbitrary types of regression. This extension will build on the intuitions obtained in Part I, in particular those pertaining to the loss of regressor ancillarity under misspecification (Section 4).

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." The history of inference under misspecification can be traced to Cox (1961, 1962), Eicker (1963), Berk(1966, 1970), Huber (1967), before being systematically elaborated by White in a series of articles (White 1980a, 1980b, 1981, 1982, among others) and capped by a monograph (White 1994). More recently, 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

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	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
`MedianInc ($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

TABLE 1
LA Homeless Data: Comparison of Standard Errors.

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 examine homelessness in relation to covariates for demographics and building usage (Berk et al. 2008). 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: $\boldsymbol{SE}_{lin}$ from linear models theory, $\boldsymbol{SE}_{boot}$ from the x-y bootstrap ($N_{boot} = 100,000$) and $\boldsymbol{SE}_{sand}$ 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 $\boldsymbol{SE}_{sand}/\boldsymbol{SE}_{boot}$ show that the sandwich and bootstrap estimators are in good agreement. Not so for the linear models estimates: we have $\boldsymbol{SE}_{boot}, \boldsymbol{SE}_{sand} > \boldsymbol{SE}_{lin}$ for the regressors `PercVacant`, `PercCommercial` and `PercIndustrial`, and $\boldsymbol{SE}_{boot}, \boldsymbol{SE}_{sand} < \boldsymbol{SE}_{lin}$ for `Intercept`, `MedianInc ($1000)`, `PercResidential`. Only for `PercMinority` is $\boldsymbol{SE}_{lin}$ off by less than 10% from $\boldsymbol{SE}_{boot}$ and $\boldsymbol{SE}_{sand}$. The discrepancies affect outcomes of some of the t-tests: Under linear models theory the regressors `PercCommercial` and `PercIndustrial` have commanding 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 `MedianInc ($K)` the t-value -0.977 from linear models theory becomes borderline significant with the bootstrap or sandwich estimator if the plausible one-sided alternative with negative sign is used.

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

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

3. THE POPULATION FRAMEWORK FOR LINEAR OLS

Model-robust inference means that either no working model is assumed or that the working model is not necessarily assumed to be correct. This raises the question what the meaning of parameters is. To answer this question we first introduce notation for data distributions that are free of model assumptions, essentially relying on iid sampling of x-y tuples; subsequently we introduce OLS parameters as statistical functionals of these distributions.

3.1 Populations for Linear OLS Regression

In an assumption-lean, model-robust population framework for linear OLS regression with random regressors, the ingredients are *regressor random variables* X_1, ..., 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

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

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

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

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

for the *column random vector* of the regressor variables, and $\vec{\boldsymbol{x}} = (1, x_1, ..., x_p)'$ 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 Linear OLS Statistical Functional

We write any function $f(X_1, ..., X_p)$ 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 $\vec{\boldsymbol{X}}$:

$$\mu(\vec{\boldsymbol{X}}) \;:=\; \boldsymbol{E}[\,Y\,|\,\vec{\boldsymbol{X}}\,] \;=\; \mathrm{argmin}_{f(\vec{\boldsymbol{X}})\in L_2(\boldsymbol{P})}\, \boldsymbol{E}[(Y - f(\vec{\boldsymbol{X}}))^2]\,.$$

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

$$\boxed{\begin{aligned} \boldsymbol{\beta}(\boldsymbol{P}) \;&:= \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \boldsymbol{E}[(Y - \boldsymbol{\beta}'\,\vec{\boldsymbol{X}})^2] \quad &&= \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}Y\,] \\ &= \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \boldsymbol{E}[(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\,\vec{\boldsymbol{X}})^2] &&= \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})\,]. \end{aligned}}$$

Both right hand expressions follow from the population normal equations:

$$\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\,\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}Y] \;=\; \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\,\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})] \;=\; \boldsymbol{0}. \tag{4}$$

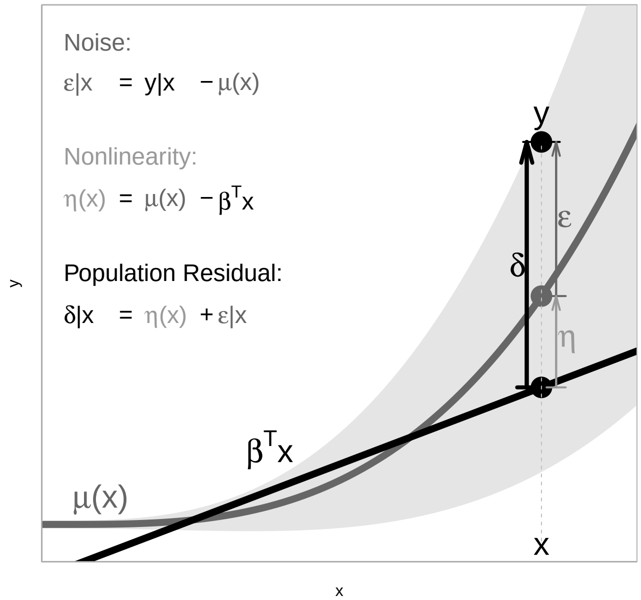

Fig 1. *Illustration of the decomposition* (5) *for linear OLS.*

The population coefficients $\boldsymbol{\beta}(\boldsymbol{P}) = (\beta_0(\boldsymbol{P}), \beta_1(\boldsymbol{P}), ..., \beta_p(\boldsymbol{P}))'$ 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 of course $\boldsymbol{\beta}(\boldsymbol{P}) = \tilde{\boldsymbol{\beta}}$. 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.

3.3 The Noise-Nonlinearity Decomposition for Population OLS

The response Y has the following canonical decompositions:

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

We call ϵ the *noise* and η the *nonlinearity*, while for δ there is no standard term, but "*population residual*" may suffice; see Table 2 and Figure 1. Important to note is that (5) is a decomposition; it makes no model assumptions on δ or ϵ. 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 $\eta(\vec{\boldsymbol{X}})$, and a random component, the noise ϵ. In model-trusting linear modeling, one assumes $\eta(\vec{\boldsymbol{X}}) \stackrel{\boldsymbol{P}}{=} 0$ and ϵ to have the same $\vec{\boldsymbol{X}}$-conditional distribution in all of regressor space, that is, ϵ is assumed

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

TABLE 2
Random variables and their canonical decompositions.

independent of $\vec{\boldsymbol{X}}$ if the latter is treated as random. No such assumptions are made here. What is left are orthogonalities satisfied by η and ϵ in relation to $\vec{\boldsymbol{X}}$. If we call independence "strong-sense orthogonality", we have instead

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

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

In what follows we will need some natural definitions:

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

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

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

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

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

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

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

4. CONSPIRACY I: NONLINEARITY AND RANDOM X JOINTLY AFFECT SLOPE PARAMETERS

4.1 Nonlinearity Destroys Regressor Ancillarity for Slopes

Conditioning on the regressors and hence treating them as fixed when in fact they are random has historically been justified with the ancillarity principle.

Regressor ancillarity is a property of working models $p(y \,|\, \vec{\boldsymbol{x}};\, \boldsymbol{\theta})$ for the conditional distribution of $Y|\vec{\boldsymbol{X}}$, where $\boldsymbol{\theta}$ is the parameter of interest in the traditional meaning of a parametric model. Because we treat $\vec{\boldsymbol{X}}$ as random, the assumed joint distribution of $(Y, \vec{\boldsymbol{X}})$ is

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

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

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

which are free of $p(\vec{\boldsymbol{x}})$, detaching the regressor distribution from the parameter $\boldsymbol{\theta}$. (For further discussion of ancillarity see Appendix B.) This logic is valid if $p(y \,|\, \vec{\boldsymbol{x}};\, \boldsymbol{\theta})$ correctly describes the actual conditional regressor distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ for some $\boldsymbol{\theta}$. If, however, there is no $p(y \,|\, \vec{\boldsymbol{x}};\, \boldsymbol{\theta})$ that describes $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ correctly, then regressor ancillarity is lost and the regressor distribution becomes intertwined with the parameters. In order to pursue the consequences of the loss of regressor ancillarity, one needs to step outside the working model and interpret parameters as statistical functionals following Section 3. The proposition below describes the consequences of broken regressor ancillarity for the statistical functional $\boldsymbol{\beta}(\boldsymbol{P})$ if the conditional mean function $\mu(\vec{\boldsymbol{X}})$ is *not* linear:

Proposition 4.1: Breaking Regressor Ancillarity in linear OLS
Consider joint distributions $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ that share a function $\mu(\vec{\boldsymbol{x}})$ as their conditional expectation of the response. Among them, the functional $\boldsymbol{\beta}(\boldsymbol{P})$ will depend on the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ if and only if $\mu(\vec{\boldsymbol{x}})$ is nonlinear.

[This is a loose statement; for more precision, see Appendix D.1.] 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. [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, $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$. What may seem to be superficially contradictory inferences from the two studies may be compatible if (1) the true response surface $\mu(\vec{\boldsymbol{x}})$ is not linear and (2) the regressors' high-density regions differ between studies. Differences in estimated parameter values often become

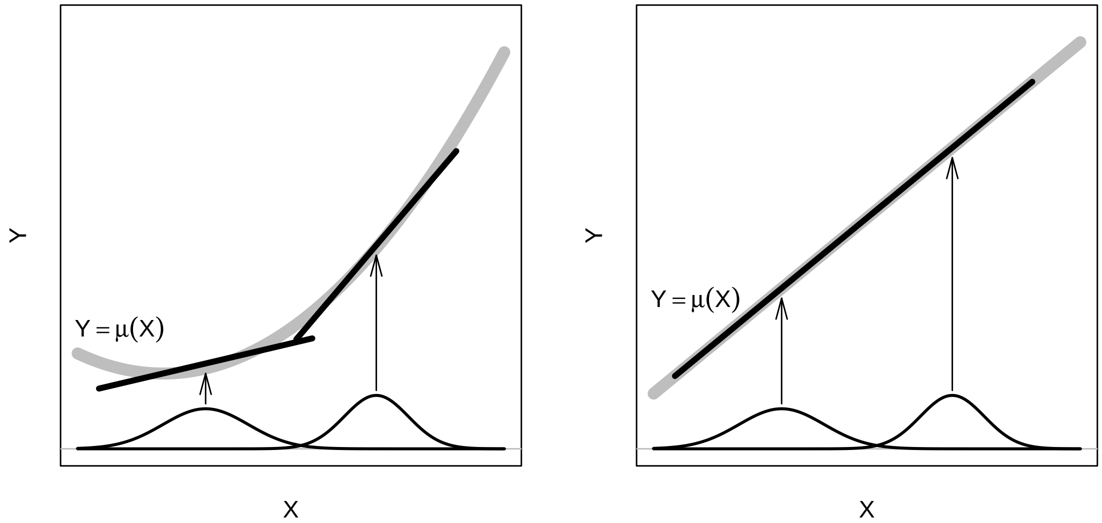

FIG 2. *Illustration of the dependence of the population OLS solution on the marginal distribution of the regressors: The left figure shows dependence in the presence of nonlinearity; the right figure shows independence in the presence of linearity.*

visible in meta-analyses and are labeled "parameter heterogeneity." The source of this heterogeneity may be differences in regressor distributions combined with model misspecification. — The single-regressor situation of Figure 2 gives only an insufficient impression of the true difficulties arising from differences in regressor distributions. While such differences are easily detected for one regressor, they can become increasingly complex even in moderate regressor dimensions and virtually undiagnosable in their effects on the parameters across studies if only marginal descriptive statistics are reported.

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 chance of appearing "well-specified" because approximations work better over narrow ranges. In the figure the narrow range of the regressor distribution $\boldsymbol{P}_1(d\vec{\boldsymbol{x}})$ is the reason why the linear approximation is very nearly well-specified, whereas the wide range of $\boldsymbol{P}_2(d\vec{\boldsymbol{x}})$ is the reason for the gross misspecification of the linear approximation. — Again, the situation gets increasingly complicated in higher regressor dimensions where the notion of "regressor range" takes on a multivariate meaning.

5. OLS ESTIMATION, CONDITIONAL PARAMETERS, AND THE ASSOCIATED NOISE-NONLINEARITY DEPOMPOSITION

We turn from populations to estimation from iid data. With the focus on iid sampling we sacrifice the generality found, for example, in White (1980b, 1994) and Hansen (1982), but the greater simplicity has didactic advantages.

We denote iid observations from a joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ by $(Y_i, \vec{\boldsymbol{X}}_i{}') = (Y_i, 1, X_{i,1}, ..., X_{i,p})$ $(i = 1, 2, ..., N)$. We stack them to vectors and matrices as in Table 3, prepending a constant 1 to the regressors to allow an intercept term. In particular, $\vec{\boldsymbol{X}}_i{}'$ is the i'th row and $\mathbf{X}_j$ the j'th column of the regressor

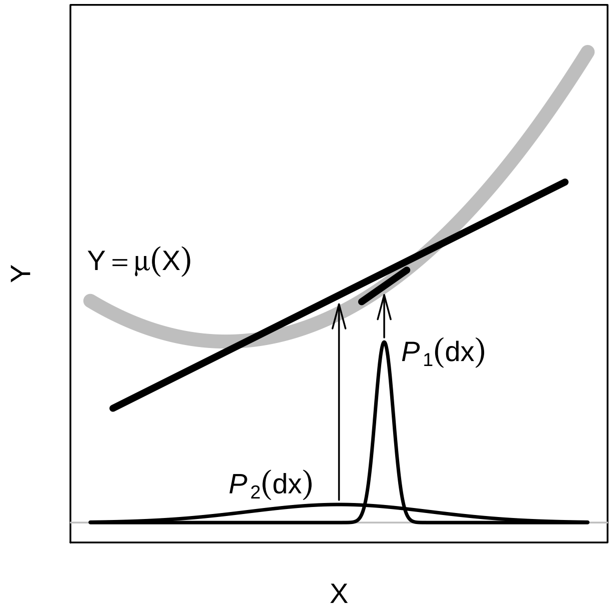

Fig 3. *Illustration of the interplay between regressors' high-density range and nonlinearity: Over the small range of $\boldsymbol{P}_1$ the nonlinearity will be undetectable and immaterial for realistic sample sizes, whereas over the extended range of $\boldsymbol{P}_2$ the nonlinearity is more likely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

matrix $\mathbf{X}$ $(i = 1, ..., N,\ j = 0, ..., p)$.

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

$$\text{(9)} \qquad \boldsymbol{\eta} \ = \ \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} \ = \ \mathbf{Y} - \boldsymbol{\mu}, \qquad \boldsymbol{\delta} \ = \ \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \ = \ \boldsymbol{\eta} + \boldsymbol{\epsilon}.$$

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

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

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

$$\text{(10)} \qquad \hat{\boldsymbol{\beta}} \ = \ \mathrm{argmin}_{\tilde{\boldsymbol{\beta}}} \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\|^2 \ = \ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{Y}\,.$$

Because we are not conditioning on $\mathbf{X}$, randomness of $\hat{\boldsymbol{\beta}}$ stems from $\mathbf{Y}$ as well as $\mathbf{X}$. The sample residual vector $\boldsymbol{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, which arises from $\hat{\boldsymbol{\beta}}$, is distinct from the population residual vector $\boldsymbol{\delta} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, which arises from $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$. If we write $\hat{\boldsymbol{P}}$ for the empirical distribution of the N observations $(Y_i, \vec{\boldsymbol{X}}_i{}')$, then $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$ is the plug-in estimate. — In $\mathbf{X}$-conditional or fixed-$\mathbf{X}$ theory the target of estimation $\boldsymbol{\beta}(\mathbf{X})$ is what we may call the "*conditional parameter*":

$$\boxed{\boldsymbol{\beta}(\mathbf{X}) \ := \ \mathrm{argmin}_{\boldsymbol{\beta}}\, \boldsymbol{E}[\,\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \,|\, \mathbf{X}\,] \ = \ (\mathbf{X}'\,\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{\mu} \ = \ \boldsymbol{E}[\,\hat{\boldsymbol{\beta}} \,|\, \mathbf{X}\,].}$$

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

TABLE 3

Random variable notation for estimation in linear OLS based on iid observational data.

In random-$\mathbf{X}$ theory, on the other hand, the target of estimation is $\boldsymbol{\beta}(\boldsymbol{P})$, while the 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 natural decomposition:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\boldsymbol{P}) \;=\; (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\mathbf{X})) \;+\; (\boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P})). \tag{11}$$

This in turn corresponds to the decomposition $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$:

Definition and Lemma 5: *Define "Estimation Offsets" (EOs) as follows:*

$$\boxed{\begin{array}{llcl} \textit{Total EO} & := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\boldsymbol{P}) & = & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\delta}, \\ \textit{Noise EO} & := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\mathbf{X}) & = & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}, \\ \textit{Approximation EO} & := \boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P}) & = & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta}. \end{array}} \tag{12}$$

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

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,\vert\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}, \quad \boldsymbol{\beta}(\boldsymbol{P}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}).$$

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

6. CONSPIRACY II: NONLINEARITY AND RANDOM X JOINTLY CREATE SAMPLING VARIATION

6.1 Sampling Variance Canonically Decomposed

For the conditional parameter $\boldsymbol{\beta}(\mathbf{X})$ to be a non-trivial random variable, two factors need to be present: (1) the regressors need to be random and (2) the

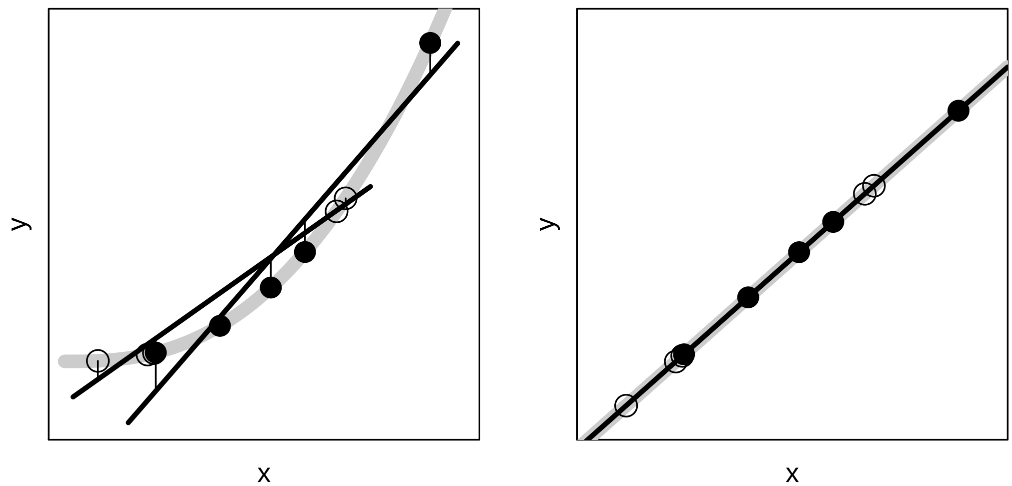

FIG 4. *Noise-less Response: The filled and the open circles represent two "datasets" from the same population. The x-values are random; the y-values are a deterministic function of x: $y = \mu(x)$ (shown in gray).*
Left: The true response $\mu(x)$ is nonlinear; the open and the filled circles have different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

nonlinearity must not (a.s.) vanish. In combination, they produce the second "conspiracy" in the title of the article. This is most easily seen from the form of the approximation EO in (12) which depends on the random matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the vector of nonlinearities $\boldsymbol{\eta}$. The full, unconditional variability of $\hat{\boldsymbol{\beta}}$ is then no longer solely due to the conditional distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ of the response Y; it contains a contribution owed to the marginal distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ of the regressors $\vec{\boldsymbol{X}}$. This fact is reflected by the following natural decomposition:

$$\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \boldsymbol{E}[\,\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\mathbf{X}]] \;+\; \boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\mathbf{X}]]\,, \tag{13}$$

where the left hand side represents the full variability of $\hat{\boldsymbol{\beta}}$, including the variability due to $\mathbf{X}$. In view of Lemma 5 this decomposition corresponds to $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$:

$$\begin{aligned}
\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,] \;&=\; \boldsymbol{V}[\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{\delta}\,]\,,\\
\boldsymbol{E}[\,\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\mathbf{X}]] \;&=\; \boldsymbol{E}[\,\boldsymbol{V}[\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{\epsilon}\,|\mathbf{X}]\,]\\
&\boxed{\boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\mathbf{X}]\,] \;=\; \boldsymbol{V}[\,\boldsymbol{\beta}(\mathbf{X})\,] \;=\; \boldsymbol{V}[\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{\eta}\,]}
\end{aligned} \tag{14}$$

Shown in the box is the "conspiracy" contribution to the sampling variance due to the conditional parameter $\boldsymbol{\beta}(\mathbf{X})$ which represents a joint effect of the vector of nonlinearities $\boldsymbol{\eta}$ and the randomness of $\mathbf{X}$.

6.2 Sampling Variability Due to the Conspiracy Illustrated

The "conspiracy" of random $\mathbf{X}$ and $\boldsymbol{\eta}$ is best illustrated graphically, similar to Section 4. 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 $Y = \mu(\vec{\boldsymbol{X}})$ where $\mu(\cdot)$ is some non-linear function (that is, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} = \delta_{\mu(\vec{\boldsymbol{X}})}$

are point masses). In the decomposition of Section 6.1 the noise term $\boldsymbol{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \mathbf{0}$ vanishes a.s.

A graphical depiction 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 exhibit sampling variability due to the nonlinearity of the response. This effect is absent for a linear response shown in the right hand frame. — A more dramatic illustration is available in the animation whose URL is given in the introduction.

6.3 The Quandary of Fixed-X Theory and the Need for Random-X Theory

The fixed-$\mathbf{X}$ approach of linear models theory necessarily assumes well-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 other source, the approximation EO $\boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}(\boldsymbol{P})$ arising from the regressor distribution interacting with nonlinearity. The remedy of fixed-$\mathbf{X}$ theory is to call for model diagnostics and declare a model and its inferences to be invalid if misspecification is detected.

While model diagnostics should be mandatory in any data analysis, rejecting inferences based on diagnostics could sometimes be avoided: There do exist misspecifications that do not invalidate the standard errors of linear models theory; see Section 11.6 for examples. Section 12 proposes a test to detect which slopes have their usual standard errors invalidated by misspecification.

Good data analysts will of course not be defeated by model diagnostics and instead continue with data-driven modeling, for example, by adding terms to the fitted equation and growing the column dimension of $\mathbf{X}$. This can be done using model selection based on formal algorithms and/or successive informal residual diagnostics. Such model selection, however, invalidates both classical statistical inference and residual diagnostics. While recent work addresses the problem of post-selection inference (e.g., Berk et al. 2013; Lee et al. 2016), we would not know how to address what we may call "Mammen's dilemma": Mammen's (1996) results imply for models with numerous regressors that, roughly speaking, residual distributions look as assumed by the working model (e.g., Gaussian for OLS), irrespective of the true error distribution. Therefore, as much as diagnostics should be mandatory in every data analysis, there are inherent limits to what they can achieve. Satisfactory diagnostics may yet hide misspecification.

When misspecification is ignored in linear OLS, then both nonlinearity and heteroskedastic noise are mistakenly treated as exchangeable noise. This mistreatment is exhibited rather visibly in the residual bootstrap. Asymptotically correct treatment is, however, provided by (1) sandwich estimators resulting from asymptotic plug-in in model-robust random-$\mathbf{X}$ theory, and (2) by x-y bootstrap estimators of standard error. These approaches are asymptotically valid even in noise-free deterministic but misspecified situations (for fixed p and growing N). The justifications derive from central limit theorems 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 only in decomposing the overall asymptotic variance into contributions stemming from the noise EO and the approximation EO according to (12), thereby providing an asymptotic analog of the finite-sample decomposition of sampling variance in Section 6.1.

Proposition 7: *For linear OLS the three EOs follow CLTs:*

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

These three statements once 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:

$$
(16)\qquad \boldsymbol{E}[\, m^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,] \;=\; \boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,], \qquad \boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,] \;=\; \boldsymbol{E}[\,\epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,].
$$

The asymptotic variance of linear OLS can therefore be written as

$$
(17)\qquad \boldsymbol{AV}[\boldsymbol{\beta}, \boldsymbol{P}] \;:=\; \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\, \boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\, \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}.
$$

The symbol $\boldsymbol{\beta}$ stands for the statistical functional $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ and by implication its 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; see Section 8.1.

Special cases covered by the above proposition are the following:

- **First order well-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 well-specification:** $\eta(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$, $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=}$ const. The *non*-sandwich form is asymptotically valid without Gaussian errors.

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. A closer connection between them will be established here.

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 $\boldsymbol{AV}[\boldsymbol{\beta}, \boldsymbol{P}]$ in (17) is given explicitly and also suitably continuous in the two arguments, one obtains a consistent estimator by plugging in $\hat{\boldsymbol{P}}$ for $\boldsymbol{P}$:

$$
(18)\qquad \hat{\boldsymbol{AV}}[\,\boldsymbol{\beta}\,] \;:=\; \boldsymbol{AV}[\boldsymbol{\beta}, \hat{\boldsymbol{P}}], \qquad \hat{\boldsymbol{SE}}[\beta_j] \;:=\; \frac{1}{N^{1/2}}(\hat{\boldsymbol{AV}}[\,\boldsymbol{\beta}\,])_{jj}^{1/2}.
$$

[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 $\boldsymbol{E}[...]$ with sample means $\hat{\boldsymbol{E}}[...]$, $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ 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

$$\hat{\boldsymbol{E}}[\, r^2 \, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}' \,] = \tfrac{1}{N}\,(\mathbf{X}'\,\boldsymbol{D}_{\boldsymbol{r}}^2\,\mathbf{X}), \qquad \hat{\boldsymbol{E}}[\, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}' \,] = \tfrac{1}{N}\,(\mathbf{X}'\mathbf{X}).$$

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

$$\begin{aligned} \hat{\boldsymbol{AV}}_{sand}[\,\boldsymbol{\beta}\,] \;&:=\; \hat{\boldsymbol{E}}[\, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}' \,]^{-1}\,\hat{\boldsymbol{E}}[\, r^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}' \,]\;\hat{\boldsymbol{E}}[\, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}' \,]^{-1} \\ &=\; N\,(\mathbf{X}'\mathbf{X})^{-1}\,(\mathbf{X}'\,\boldsymbol{D}_{\boldsymbol{r}}^2\,\mathbf{X})\,(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{19}$$

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 (18) 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 leave-one-out residuals. The following subsection will rely on the original version (19).

8.2 The M-of-N Bootstrap Estimator of Asymptotic Variance

An alternative to plug-in is estimating asymptotic variance with the x-y bootstrap. To link plug-in and bootstrap estimators we need the M-of-N bootstrap where the *resample size* M may differ from the sample size N. 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 $M < N$. 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. Ours is a well behaved context, hence there is no need for $M \ll N$; instead, we consider bootstrap resampling for the extreme case $M \gg N$, namely, the limit $M \to \infty$.

The crucial observation is as follows: Because bootstrap resampling is iid sampling from some distribution, there holds a CLT as the resample size grows, $M \to \infty$, holding N fixed. It is immaterial that, in this case, the sampled distribution is the empirical distribution $\hat{\boldsymbol{P}} = \hat{\boldsymbol{P}}_N$ of a given dataset $\{(Y_i, \vec{\boldsymbol{X}}_i{}')\}_{i=1...N}$, which is frozen of size N as $M \to \infty$. To execute this idea, we adapt the first CLT of Section 7 to bootstrap estimates $\boldsymbol{\beta}_M^* = \boldsymbol{\beta}(\boldsymbol{P}_M^*)$, where $\boldsymbol{P}_M^*$ is the empirical distribution of bootstrap data $\{(Y_i^*, \vec{\boldsymbol{X}}_i^*{}')\}_{i=1...M}$ drawn iid from $\hat{\boldsymbol{P}}_N$. In detail, make the following substitutions, subscripting all objects with their respective sample or resample sizes: $\hat{\boldsymbol{\beta}}_N \mapsto \boldsymbol{\beta}_M^*$, $\boldsymbol{\beta} \mapsto \hat{\boldsymbol{\beta}}_N$, $\boldsymbol{P} \mapsto \hat{\boldsymbol{P}}_N$ and $\boldsymbol{AV}[\,\boldsymbol{\beta}, \boldsymbol{P}\,] \mapsto \hat{\boldsymbol{AV}}_N\,[\,\boldsymbol{\beta}\,]$. As a result we obtain the following observation:

Proposition 8.2:

$$M^{1/2}\,(\boldsymbol{\beta}_M^* - \hat{\boldsymbol{\beta}}_N) \;\; \xrightarrow{\mathcal{D}} \;\; \mathcal{N}\left(\mathbf{0},\; \hat{\boldsymbol{AV}}_N\,[\,\boldsymbol{\beta}\,]\right) \qquad (M \to \infty, \quad N \textit{ fixed}). \tag{20}$$

Corollary 8.2: *The sandwich estimator* (19) *for OLS slope estimators is the asymptotic variance estimated by the M-of-N bootstrap in the limit $M \to \infty$ for a fixed sample of size N.*

This fact provides a natural link between the two apparently independent approaches to model-robust standard errors. Sandwich estimators have the advantage that they result in unique standard error values whereas bootstrap standard errors have simulation error in practice. On the other hand, the x-y bootstrap is more flexible because the bootstrap distribution can be used to generate confidence intervals that are second order correct (see, e.g., Efron and Tibshirani 1994; Hall 1992; McCarthy, Zhang et. al. 2016).

For further connections see MacKinnon and White (1985): 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. A deep connection between jackknife and bootstrap is given by Wu (1986).

9. ADJUSTED REGRESSORS

This section prepares the ground for two projects: (1) proposing meanings of slopes in the presense 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 C 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 $X_{j\bullet}$:

$$\beta_j \;=\; \frac{\boldsymbol{E}[\,Y_{\bullet -j} X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}^2]} \;=\; \frac{\boldsymbol{E}[\,Y X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}^2]} \;=\; \frac{\boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}}) X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}^2]}\,. \tag{21}$$

  The rightmost 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 regressors. 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 $\mathbf{X}_{j\bullet}$:

$$\hat{\beta}_j \;=\; \frac{\langle \mathbf{Y}_{\hat{\bullet} -j}, \mathbf{X}_{j\hat{\bullet}} \rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^2} \;=\; \frac{\langle \mathbf{Y}, \mathbf{X}_{j\hat{\bullet}} \rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^2}\,. \tag{22}$$

[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. THE MEANING OF SLOPES IN THE PRESENCE OF NONLINEARITY

A first use of regressor adjustment is for proposing a meaning of linear slopes in the presence of nonlinearity, and thereby 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 hold that the parameter is not intrinsically wrong, rather, it is in need of a useful interpretation.

The issue, on which everyone agrees, 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. The challenge is to provide an alternative interpretation that remains valid and intuitive. As mentioned, a plausible approach is to use adjusted variables, in which case it is sufficient to solve the interpretation problem for simple regression through the origin. In a sense to be made precise, regression slopes can then be interpreted as weighted averages of "case-wise" and "pairwise" slopes. This interpretation holds even for regressors that are nonlinearly related, as in $X_2 = X_1^2$ or $X_3 = X_1 X_2$, because the clause "at fixed levels of all other regressors" is replaced by reference to "(linearly) adjusted regressors."

To lighten the notational burden, we drop subscripts from adjusted variables:

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

By (21) and (22), the population slopes and their estimates are, respectively,

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

Slope interpretation will be based on the following devices:

- **Population parameters** β can be represented as weighted averages of ...
  - **case-wise slopes**:

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

    hence for a random case (x, y), b is its case-wise slope and w is its case-wise weight.
  - **pairwise slopes**:

    $$
    \beta \;=\; \boldsymbol{E}[\,w\,b\,], \quad \text{where} \quad b \;:=\; \frac{y - y'}{x - x'}, \quad w \;:=\; \frac{(x - x')^2}{\boldsymbol{E}[\,(x - x')^2\,]},
    $$

    hence for two independent identically distributed random cases (x, y) and (x', y'), b is their pairwise slope and w their pairwise weight.

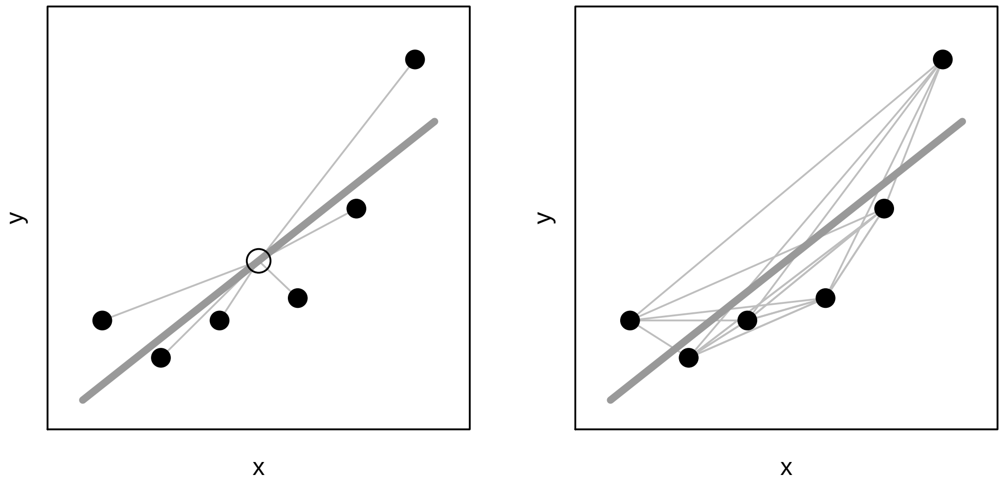

FIG 5. *Case-wise and pairwise average weighted slopes illustrated: Both plots show the same six points ("cases") as well as the OLS line fitted to them (fat gray). The left hand plot shows the case-wise slopes from the mean point (open circle) to the six cases, while the right hand plot shows the pairwise 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.*

- **Sample estimates** $\hat{\beta}$ can be represented as weighted averages of ...
  - **case-wise slopes**:

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

    hence b_i are case-wise slopes and w_i are case-wise weights.
  - **pairwise slopes**:

$$\hat{\beta} = \sum_{ik} w_{ik} \, b_{ik} \,, \quad \text{where} \quad b_{ik} := \frac{y_i - y_k}{x_i - x_k}, \quad w_{ik} := \frac{(x_i - x_k)^2}{\sum_{i'k'} (x_{i'} - x_{k'})^2},$$

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

See Figure 5 for an illustration for samples. — In the LA homeless data, we can interpret the slope for the regressor `PercVacant`, say, in the following ways:

(1) "Adjusted for all other regressors, the mean deviation of `Homeless` in relation to the mean deviation of `PercVacant` is estimated to average between 4 and 5 homeless per one percent of vacant property."

(2) "Adjusted for all other regressors, the difference in `Homeless` between two census tracts in relation to their difference in `PercVacant` is estimated to average between 4 and 5 homeless per one percent of vacant property."

Missing is a technical reference to the fact that the "average" is weighted. All such formulations, if they aspire to be technically correct, end up being inelegant, but the same is the case with the model-trusting formulation:

(*) "At constant levels of all other regressors, the average difference in `Homeless` for a one percent difference in `PercVacant` is estimated to be between 4 and 5 homeless."

This statement is strangely abstract as it refers to an unreal mental scenario of pairs of census tracts that agree in all other regressors but differ in the focal regressor by one unit. By comparison, statements (1) and (2) above refer to observed mean deviations and differences. In practice, users will run with the shorthand "the slope for `PercVacant` is between 4 and 5 homeless per one percent."

The formulas support the intuition that, even in the presence of nonlinearity, a linear fit can be used to infer the overall direction of the association between the response and a regressor, adjusted for all other regressors. It is of course possible to construct examples where no single direction of association exists, as when $\boldsymbol{E}[Y|X] = \mu(X) = X^2$ and X is distributed symmetrically about 0. If, however, $X > 0$ a.s., then the direction of association is certainly positive, pointing yet again to the crucial role of the regressor distribution in the presence of nonlinearity; moreover, if $|\boldsymbol{E}[X]|/\boldsymbol{SD}[X] \gg 0$, a linear fit provides an excellent approximation to $\mu(X) = X^2$.

Finally, there is precedent in thinking that the interpretation of OLS slopes as averages of observed slopes is natural and accessible to large audiences: The above formulas were used by Gelman and Park (2008) with the "Goal of Expressing Regressions as Comparisons that can be Understood by the General Reader" (see their Sections 1.2 and 2.2). The formulas also have a considerable pedigree: Stigler (2001) includes Edgeworth in their long history, while Berman (1988) traces them back to a 1841 article by Jacobi written in Latin. A powerful generalization based on tuples rather than pairs of $(y_i, \vec{\boldsymbol{x}}_i')$ rows was used by Wu (1986) for the analysis of jackknife and bootstrap procedures (see his Section 3, Theorem 1).

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 (11) can be written componentwise using adjustment as follows:

$$\boxed{\begin{aligned} &\textit{Total EO}: & \hat{\beta}_j - \beta_j &= \frac{\langle \mathbf{X}_{j\bullet}, \boldsymbol{\delta} \rangle}{\|\mathbf{X}_{j\bullet}\|^2}, \\ &\textit{Noise EO}: & \hat{\beta}_j - \beta_j(\mathbf{X}) &= \frac{\langle \mathbf{X}_{j\bullet}, \boldsymbol{\epsilon} \rangle}{\|\mathbf{X}_{j\bullet}\|^2}, \\ &\textit{Approximation EO}: & \beta_j(\mathbf{X}) - \beta_j &= \frac{\langle \mathbf{X}_{j\bullet}, \boldsymbol{\eta} \rangle}{\|\mathbf{X}_{j\bullet}\|^2}. \end{aligned}} \tag{23}$$

To see these identities directly, note the following, in addition to (22): $\boldsymbol{E}[\hat{\beta}_j|\mathbf{X}] = \langle \boldsymbol{\mu}, \mathbf{X}_{j\bullet} \rangle / \|\mathbf{X}_{j\bullet}\|^2$ and $\beta_j = \langle \mathbf{X}\boldsymbol{\beta}, \mathbf{X}_{j\bullet} \rangle / \|\mathbf{X}_{j\bullet}\|^2$, the latter due to $\langle \mathbf{X}_{j\bullet}, \mathbf{X}_k \rangle = \delta_{jk} \|\mathbf{X}_{j\bullet}\|^2$. Finally use $\boldsymbol{\delta} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}$. $\square$

From (23), asymptotic normality of the coefficient-specific EOs can be separately expressed using population adjustment:

Corollary 11.1:

$$\boxed{\begin{array}{lcl} N^{1/2}(\hat{\beta}_j - \beta_j) & \xrightarrow{\mathcal{D}} & \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\, m^2(\vec{\boldsymbol{X}}) X_{j\bullet}^2]}{\boldsymbol{E}[\, X_{j\bullet}^2]^2}\right) \;=\; \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\, \delta^2 X_{j\bullet}^2]}{\boldsymbol{E}[\, X_{j\bullet}^2]^2}\right) \\[3ex] N^{1/2}(\hat{\beta}_j - \beta_j(\mathbf{X})) & \xrightarrow{\mathcal{D}} & \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\, \sigma^2(\vec{\boldsymbol{X}}) X_{j\bullet}^2]}{\boldsymbol{E}[\, X_{j\bullet}^2]^2}\right) \;=\; \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\, \epsilon^2 X_{j\bullet}^2]}{\boldsymbol{E}[\, X_{j\bullet}^2]^2}\right) \\[3ex] N^{1/2}(\beta_j(\mathbf{X}) - \beta_j) & \xrightarrow{\mathcal{D}} & \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\, \eta^2(\vec{\boldsymbol{X}}) X_{j\bullet}^2]}{\boldsymbol{E}[\, X_{j\bullet}^2]^2}\right) \end{array}}$$

The equalities on the right side in the first and second case are based on (16). 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 of Corollary 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 11.2: *Proper Asymptotic Variance.*

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

Reflecting Corollary 11.1, the conditional MSE decomposition $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ (8) translates to

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

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 (Subsection 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, $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(N - p - 1)$:

$$\hat{\sigma}^2 \;\xrightarrow{\boldsymbol{P}}\; \boldsymbol{E}[\, \delta^2\,] \;=\; \boldsymbol{E}[\, m^2(\vec{\boldsymbol{X}})\,] \;=\; \boldsymbol{E}[\, \sigma^2(\vec{\boldsymbol{X}})\,] + \boldsymbol{E}[\, \eta^2(\vec{\boldsymbol{X}})\,], \qquad N \to \infty.$$

Thus, the model-robust limit of $\hat{\sigma}^2$ is the average conditional MSE of Y, which decomposes into the sum of (1) the average conditional noise variance and (2) the average squared nonlinearity.

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

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \hat{\sigma}^2\,(\mathbf{X}'\mathbf{X})^{-1}, \qquad \hat{\boldsymbol{SE}}^2_{lin}[\,\hat{\beta}\,] \;=\; \frac{\hat{\sigma}^2}{\|\mathbf{X}_{j\hat{\bullet}}\|^2}. \tag{24}$$

Their scaled limits under model-robust assumptions are as follows:

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

These limits are valid in a model-robust sense, but they are **model-trusting** asymptotic variances that provide valid standard errors if the first and second order assumptions of linear models theory hold.

Definition 11.3: *Improper Asymptotic Variance.*

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

The conditional MSE decomposition $m^2(\vec{\boldsymbol{X}})=\sigma^2(\vec{\boldsymbol{X}})+\eta^2(\vec{\boldsymbol{X}})$ again translates to

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

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 11.4: *Ratio of Asymptotic Variances, Proper/Improper (**RAV**).*

$$\boxed{\boldsymbol{RAV}[\beta_j, m^2] \;\;:=\;\; \frac{\boldsymbol{AV}_{lean}[\beta_j, m^2]}{\boldsymbol{AV}_{lin}[\beta_j, m^2]} \;\;=\;\; \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})X^2_{j\bullet}]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]\;\boldsymbol{E}[X^2_{j\bullet}]}.}$$

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 D.2, $\boldsymbol{RAV}[\beta_j, m^2]$ is a weighted mixture of these two terms. — The interpretation of the $\boldsymbol{RAV}$ is as follows:

$$\text{If}\;\; \boldsymbol{RAV}[\beta_j, m^2] \left\{ \begin{array}{c} >1 \\ =1 \\ <1 \end{array} \right\}, \text{then } \hat{\boldsymbol{SE}}_{lin}[\hat{\beta}_j] \text{ is asymptotically } \left\{ \begin{array}{c} \text{too small} \\ \text{correct} \\ \text{too large} \end{array} \right\}.$$

If, for example, $\boldsymbol{RAV}[\beta_j, m^2] = 4$, then for large samples the proper standard error of $\hat{\beta}_j$ is about twice as large as the usual standard error.

If, however, $\boldsymbol{RAV}[\beta_j, m^2] = 1$, it does *not* follow that the conditional response mean is linear and/or the conditional response variance is constant. Subsection 11.6 will show examples of nonlinearities and heteroskedasticities that result in $\boldsymbol{RAV} = 1$.

We will later have use for the following sufficient condition for $\boldsymbol{RAV} = 1$. It says essentially that when the population residual δ is a traditional error term, then the usual standard error of linear models theory is asymptotically correct. The condition is equivalent to first and second order well-specification, that is, linearity and homoskedasticity.

Lemma 11.4: *If δ^2 and ${X_{j\bullet}}^2$ are independent, then $\boldsymbol{RAV}[\beta_j, m^2] = 1$.*

Proof: The numerator of $\boldsymbol{RAV}[\beta_j, m^2]$ becomes $\boldsymbol{E}[m^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2] = \boldsymbol{E}[\,\delta^2 {X_{j\bullet}}^2] = \boldsymbol{E}[\,\delta^2]\,\boldsymbol{E}[{X_{j\bullet}}^2]$ and hence cancels with the denominator terms. □

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 non-centered 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 ∞, as will be shown next.

11.5 The Range of *RAV*:

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

Definition and Lemma 11.5: *Letting*

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

we have:

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

Thus the analysis of the $\boldsymbol{RAV}$ 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 $\boldsymbol{RAV}$ over scenarios of $m^2(\vec{\boldsymbol{X}})$ or, by Lemma 11.5, of $m_j^2(X_{j\bullet}^2)$.

Proposition 11.5: *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

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

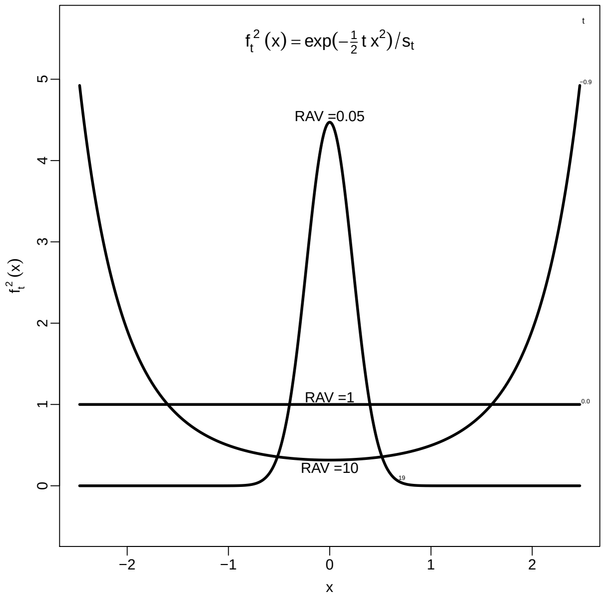

FIG 6. *A family of functions $f_t^2(x)$ that can be interpreted as conditional MSEs $m_j^2(X_{j\bullet}^2)$, heteroskedasticities $\sigma_j^2(X_{j\bullet}^2)$ or squared nonlinearities $\eta_j^2(X_{j\bullet}^2)$ (shown as functions of $x = X_{j\bullet}$ rather than $X_{j\bullet}^2$): The family interpolates* ***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 D.4 for details.)* ***RAV*** *$= \infty$ is approached as $f_t^2(x)$ bends ever more strongly in the tails of the x-distribution.* ***RAV*** *$= 0$ is approached by an ever stronger spike in the center of the x-distribution.*

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 D.3 bears this out. As the main concern is with usual standard errors that are optimistic, ***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*:

To further analyze the ***RAV***, we drill down from the conditional MSE $m_j^2(X_{j\bullet}^2)$ to conditional variance and squared nonlinearity:

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

Rather than showing curves for either case in the style of Figure 6, we translate to data scenarios in terms of heteroskedastic noise and nonlinearities: Figure 7 shows three heteroskedasticity scenarios and Figure 8 three nonlinearity scenarios. These examples train our intuitions about the types of heteroskedasticities and nonlinearities that drive the ***RAV***. According to the ***RAV*** decomposition

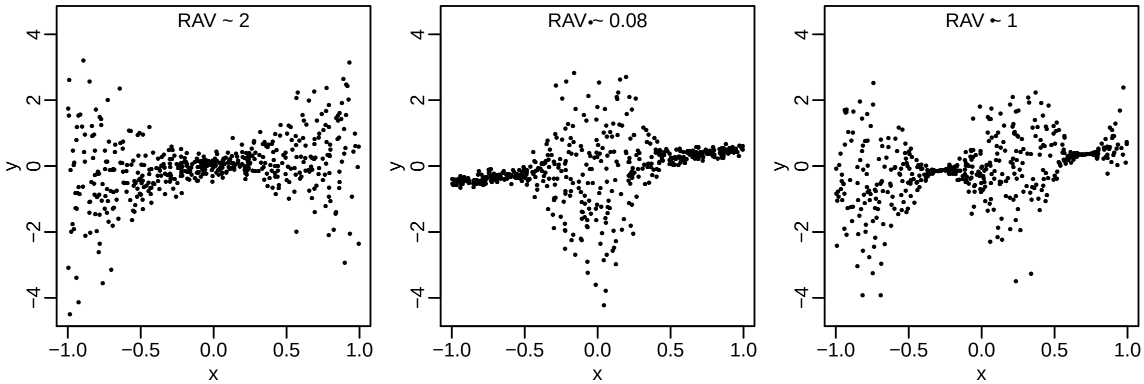

FIG 7. *The effect of heteroskedasticity on the sampling variability of slope estimates: How does the treatment of the heteroskedasticities as homoskedastic affect statistical inference?*
Left: High noise variance in the tails of the regressor distribution elevates the true sampling variability of the 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 ($\boldsymbol{RAV}[\beta_j, \sigma^2] = 1$*).*

lemma in Appendix D.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 tails 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 tails 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 $\boldsymbol{RAV} < 1$ 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 $\boldsymbol{RAV}$ decomposition lemma (Appendix D.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 considerations are heuristics for the observation that in practice $\hat{\boldsymbol{SE}}_{lin}$ is more often too small than too large compared to $\hat{\boldsymbol{SE}}_{sand}$.

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

The goal here is to write the $\boldsymbol{RAV}$ 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 proposed test is related to the class of “misspecification tests” for which there exists a literature starting with Hausman (1978) and continuing with White (1980a,b; 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 infor-

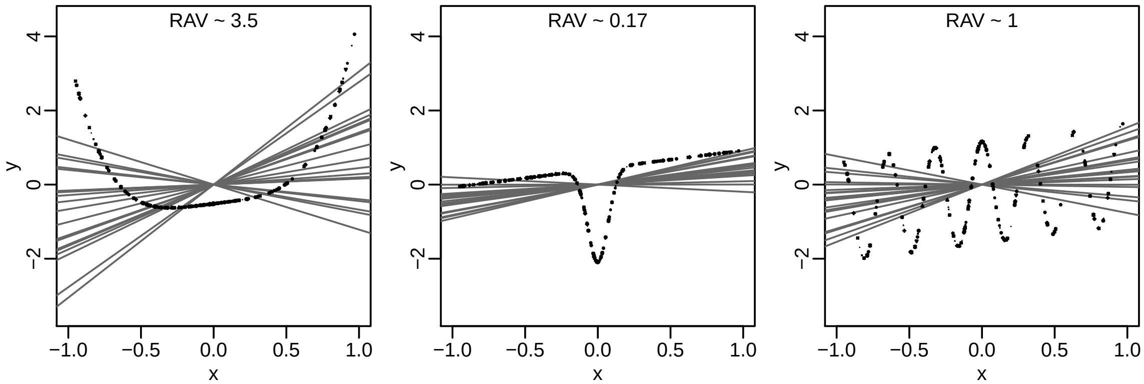

FIG 8. *The effect of nonlinearities on the sampling variability of slope estimates: The three plots show three different noise-free nonlinearities; each plot shows for one nonlinearity 20 overplotted datasets of size $N = 10$ and their fitted lines through the origin. The question is how the misinterpretation 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 ($\boldsymbol{RAV}[\beta_j, \eta^2] = 1$).

mation matrices globally, while we compare two types of standard errors, one coefficient at a time. Another, parameter-specific misspecification test of White (1982, Section 5) compares two types of coefficient estimates rather than standard error estimates, which hence is not a test of standard error discrepancies.

As illustrated above, the types of nonlinearities and heteroskedasticities that result in discrepancies between $\boldsymbol{SE}_{lin}$ and $\boldsymbol{SE}_{sand}$ are very specific ones, while other types are benign. Furthermore, different coefficients in the same model are differently affected by the same nonlinearity and heteroskedasticity because their effect on the standard errors is channeled through the adjusted regressors. The problem of standard error discrepancies is therefore not solved by general-purpose misspecification tests and model diagnostics.

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 $\boldsymbol{E}[...]$ with means $\hat{\boldsymbol{E}}[...]$, $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$, $X_{j\bullet}$ with $\mathbf{X}_{j\hat{\bullet}}$, and rescaling by N:

$$\hat{\boldsymbol{SE}}_{sand}[\hat{\beta}_j]^2 \;=\; \frac{1}{N}\frac{\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2]}{\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2]^{\,2}} \;=\; \frac{\langle \boldsymbol{r}^2, \mathbf{X}_{j\hat{\bullet}}{}^2\rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^4}. \tag{25}$$

The squaring of N-vectors is meant to be coordinate-wise. Formula (25) is algebraically equivalent to the diagonal elements of (19).

To match the raw plug-in form of the sandwich estimator (25), we use the plug-in version of the standard error estimator of linear models theory, the only difference being division by N rather than $N-p-1$:

$$\hat{\boldsymbol{SE}}_{lin}[\hat{\beta}_j]^2 \;=\; \frac{1}{N}\frac{\hat{\boldsymbol{E}}[(Y - \vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2]}{\hat{\boldsymbol{E}}[X_{j\hat{\bullet}}{}^2]} \;=\; \frac{1}{N}\frac{\|\boldsymbol{r}\|^2}{\|\mathbf{X}_{j\hat{\bullet}}\|^2}\,, \tag{26}$$

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

$$(27) \qquad \hat{\boldsymbol{RAV}}_j \;:=\; \frac{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2\,]}{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2\,]\,\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2\,]} \;=\; N\,\frac{\langle \boldsymbol{r}^2, \mathbf{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{r}\|^2\;\|\mathbf{X}_{j\hat{\bullet}}\|^2}\,.$$

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 11.4:

Proposition 12.2: *Under the null hypothesis H_0 that the population residuals δ and the adjusted regressor $X_{j\bullet}$ are independent, it holds:*

$$(28) \qquad N^{1/2}\,(\hat{\boldsymbol{RAV}}_j - 1) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0,\, \frac{\boldsymbol{E}[\,\delta^4]}{\boldsymbol{E}[\,\delta^2]^2}\,\frac{\boldsymbol{E}[X_{j\bullet}{}^4]}{\boldsymbol{E}[X_{j\bullet}^2]^2} \,-\, 1)\right).$$

As always we ignore technical assumptions. A proof outline is in Appendix D.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 D.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 non-normality test much of the time, which does not seem useful as non-normality can be diagnosed and tested by other means.

12.3 An Approximate Permutation Distribution of the *RAV* Test Statistic:

The asymptotic result of Proposition 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 non-normal 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 coordinate-wise 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

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{sand}$	$\hat{\boldsymbol{RAV}}_j$	2.5% Perm.	97.5% Perm.
(Intercept)	0.760	22.767	16.209	0.495*	0.567	3.228
MedianInc (1000)	-0.183	0.187	0.108	0.318*	0.440	5.205
PercVacant	4.629	0.901	1.363	2.071	0.476	3.852
PercMinority	0.123	0.176	0.164	0.860	0.647	2.349
PercResidential	-0.050	0.171	0.111	0.406*	0.568	3.069
PercCommercial	0.737	0.273	0.397	2.046	0.578	2.924
PercIndustrial	0.905	0.321	0.592	3.289*	0.528	3.252

TABLE 4

LA Homeless data: Permutation Inference for $\hat{\boldsymbol{RAV}}_j$ (10,000 permutations). Values of $\hat{\boldsymbol{RAV}}_j$ that fall outside the middle 95% range of their permutation null distributions are marked with asterisks. They indicate statistically significant deviations of the usual model-trusting standard errors of linear models theory from their model-robust sandwich analogs. For `MedianInc (1000)` *and* `PercResidential` *the usual standard error is too large (conservative), while for* `PercIndustrial` *it is too small (liberal). The values of approximately 2 for the $\hat{\boldsymbol{RAV}}_j$ of* `PercVacant` *and* `PercCommercial` *are not statistically significant. The $\hat{\boldsymbol{RAV}}_j$ values correspond roughly to the squares of the $\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$ values in Table 1, the minor differences stemming from using sandwich version HC2 in that table.*

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 yields an easy diagnostic of non-normality (see Appendix E for examples). Finally, by applying permutation simulations simultaneously to ***RAV*** statistics of multiple regressors, one can calibrate the retention intervals to control family-wise error. — Table 4 illustrates ***RAV*** tests with the LA Homeless data.

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. This fact suggests that use of a model-trusting standard error should be kept in mind if there is evidence in its favor, for example, through the ***RAV*** test (Section 12).

Another cost associated with the sandwich estimator is ***non-robustness in the sense of robust statistics*** (Huber and Ronchetti 2009, Hampel et al. 1986), meaning strong sensitivity to heavy-tailed distributions: The statistic $\hat{\boldsymbol{SE}}^2_{sand}[\hat{\beta}_j]$ (25) is a ratio of fourth order quantities of the data, whereas $\hat{\boldsymbol{SE}}^2_{lin}[\hat{\beta}_j]$ (26) is "only" a ratio of second order quantities. [Note we are here concerned with non-robustness of standard error estimates, not parameter estimates.] It appears that the two types of robustness are in conflict: Model-robust standard error estimators are highly non-robust to heavy tails compared to their model-trusting analogs. This is a large issue which we can only raise but not solve in this space. Here are some observations and suggestions:

- If model-robust standard errors are not classically robust, anecdotal evidence suggests that the standard errors of classical robust regression are not model-robust either. In the LA Homeless data, for example, for the most important variable `PercVacant`, we observed a ratio of 1:3.28 when comparing the standard error reported by the software (function `rlm` in the **R** *Language* (2008)) and its model-robust analog from the x-y-bootstrap.

	$\hat{\beta}_j$	SE_{lin}	SE_{boot}	SE_{sand}	$\frac{SE_{boot}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{boot}}$	t_{lin}	t_{boot}	t_{sand}
(Intercept)	2.932	0.381	0.395	0.395	1.037	1.036	0.999	7.697	7.422	7.427
MedianInc ($K)	-1.128	0.269	0.280	0.278	1.041	1.033	0.992	-4.195	-4.030	-4.061
PercVacant	1.264	0.207	0.203	0.202	0.982	0.978	0.996	6.111	6.221	6.247
PercMinority	-0.467	0.230	0.246	0.246	1.070	1.069	0.999	-2.028	-1.896	-1.897
PercResidential	-0.314	0.220	0.228	0.230	1.040	1.049	1.008	-1.432	-1.377	-1.366
PercCommercial	0.201	0.212	0.220	0.220	1.040	1.042	1.002	0.949	0.913	0.911
PercIndustrial	0.180	0.238	0.244	0.244	1.022	1.024	1.002	0.754	0.737	0.736

TABLE 5

LA Homeless Data: Comparison of Standard Errors after transforming the regressors with their cdfs to approximately uniform distributions. The taming of the tails of the regressor distributions has resolved all discrepancy issues for the usual model-trusting standard errors.

- Yet classical robust regression may confer partial robustness to the sandwich standard error as it caps residuals with a bounded ψ function. This addresses robustness to heavy tails in the vertical (y) direction.
- Robustness to outlyingness 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 D.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 iid, but the point is to examine the effect of transforming the regressors to a finite range. As a result, shown in Table 5, 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 A, Table 7), although the discrepancies are greatly reduced here, too.

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 well-specified truths. The implications of this view run deep: (1) Slope parameters need to be re-interpreted as statistical functionals arising from best-approximating linear equations to essentially arbitrary conditional mean functions; (2) the presence of nonlinearity requires new interpretations for 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; (5) nonlinearity causes slope parameters to depend not only on the conditional response distribution but on the regressor distribution as well; (6) nonlinearity causes the slope estimates to exhibit sampling variation due to the randomness of the regressors; (7) both sampling variability due to the response and due to the regressors 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 appear to be in conflict with each other.

Apart from (10), a vexing item in this list is (2): What is the meaning of a slope in the presence of nonlinearity? While we promoted an answer in terms of average observed slopes, we are aware that this issue may remain controversial. Yet, the traditional interpretation of slopes should be even more controversial: 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 well-specified. Data analysts may be of two minds about the reasonableness of assuming well-specification in some situations, but in others it may be plain that misspecification is a fact of life, as when simple models are needed for substantive reasons or for communication with consumers of statistical analysis, or when the data lend insufficient evidence about the nature of nonlinearities and/or heteroskedasticities. It may then be prudent to try approaches to interpretation and inference that do not assume well-specification.

Since White's seminal work, research into misspecification has progressed far and in many forms by addressing specific classes of misspecifications: dependencies, heteroskedasticities and nonlinearities. A direct generalization of White's sandwich estimator to time series dependence in regression data is the "heteroskedasticity and auto-correlation consistent" (HAC) estimator of standard error by Newey and West (1987). Structured second order misspecifications such as over/underdispersion have been addressed with quasi-likelihood. More generally 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 of statistics research where model-trusting theory appears frequently:

- Bayes inference, when it relies on uninformative priors, is asymptotically equivalent to model-trusting frequentist inference. It should be reasonable to ask how far inferences from Bayesian models are adversely affected by misspecification. Complex Bayesian models often use large numbers of fitted parameters and control overfitting by shrinkage, hence asymptotic comparisons may be inadequate and might have to be replaced by other forms of analysis. Some promising developments are the following: 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

appears to rely on the assumptions of linearity, homoskedasticity as well as normality of error distributions. 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). Related is also the incorporation of ideas from classical robust statistics by, for example, El Karoui et al. (2013), Donoho and Montanari (2014), and Loh (2015).

In summary, while interesting developments are in progress, there remains work to be done especially in some of today's most lively research areas. Even within the narrower, non-Bayesian and low-dimensional domain there remains the unresolved conflict between model-robustness and classical robustness at the level of standard errors. The idea that statistical models are approximations, and that this idea has consequences for statistical inference, may not yet be fully realized.

REFERENCES

[1] Aldrich, J. (2005). Fisher and Regression. *Statistical Science* **20** (4), 4001–417.

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

[3] Berk, R. H. (1966). Limiting Behavior of Posterior Distributions When the Model is Incorrect. *The Annals of Mathematical Statistics* **37** (1), 51–58.

[4] Berk, R. H. (1970). Consistency A Posteriori. *The Annals of Mathematical Statistics* **41** (3), 894–960.

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

[6] Berman, M. (1988). A Theorem of Jacobi and its Generalization. *Biometrika* **75** (4), 779–783.

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

[8] Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. in *Robustness in Statistics: Proceedings of a Workshop* (Launer, R. L., and Wilkinson, G. N., eds.) Amsterdam: Academic Press (Elsevier), 201–236.

[9] Bühlmann, P. and van de Geer, S. (2015). High-dimensional Inference in Misspecified Linear Models. **arXiv:1503.06426**

[10] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, London: Chapman & Hall.

[11] Cox, D.R. (1995). Discussion of Chatfield (1995). *Journal of the Royal Statistical Society, Series A* **158** (3), 455-456.

[12] Davies, P. L. (2014). *Data Analysis and Approximate Models.* Boca Raton, FL: CRC Press.

[13] De Blasi, P. (2013). Discussion of Walker (2013). *Journal of Statistical Planning and Inference* **143**, 1634–1637.

[14] Diggle, P. J. and Heagerty, P. and Liang, K.Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford Statistical Science Series. Oxford: Oxford University Press. ISBN 978-0-19-852484-7.

[15] Donoho, D. D. and Montanari, A. (2014). Variance Breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. **arXiv:1503.02106**

[16] Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

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

[18] Eicker, F. (1963). Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions. *The Annals of Mathematical Statistics* **34** (2), 447-456.

[19] El Karoui, N. and Bean, D. and Bickel, P. and Yu, B. (2013). Optimal M-Estimation in High-Dimensional Regression. *Proceedings of National Academy of Sciences* **110** (36), 14563-14568.

[20] Freedman, D. A. (1981). Bootstrapping Regression Models. *The Annals of Statistics* **9** (6), 1218–1228.

[21] Freedman, D. A. (2006). On the So-Called “Huber Sandwich Estimator” and “Robust Standard Errors.” *The American Statistician* **60** (4), 299–302.

[22] Gelman, A. and Park, D.. K. (2008). Splitting a Regressor at the Upper Quarter or Third and the Lower Quarter or Third, *The American Statistician* **62** (4), 1–8.

[23] Hall, A. R. (2005). *Generalized Method of Moments* (Advanced Texts in Econometrics). Oxford: Oxford University Press. ISBN 0-19-877520-2.

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

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

[26] Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **50** (4), 10291054.

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

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

[29] Hinkley, D. V. (1977). Jackknifing in Unbalanced Situations. *Technometrics* **19**, 285–292.

[30] Hoff, P. and Wakefield, J. (2013). Bayesian Sandwich Posteriors for Pseudo-True Parameters — Discussion of Walker (2013). *Journal of Statistical Planning and Inference* **143**, 1638–1642.

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

[32] Huber, P. J. and Ronchetti, E.M. (2009). *Robust Statistics.*, 2nd ed. New York, NY: Wiley.

[33] Kauermann, G. and Carroll, R. J. (2001). A Note on the Efficiency of Sandwich Covariance Matrix Estimation, *Journal of the American Statistical Association* **96**(456), 1387-1396.

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

[35] Krasker, W. S. and Welsch, R. W. (1982). Efficient Bounded-Influence Regression Estimation. *Journal of the American Statistical Association* **77** (379), 595-604.

[36] Lehmann, E. L. and Romano, J. P. (2008). *Testing Statistical Hypotheses.* New York, NY: Springer Verlag.

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

[38] Loh, P. (2015). Statistical Consistency and Asymptotic Normality for High-Dimensional Robust M-Estimators. **arXiv:1501.00312**

[39] Long, J. S. and Ervin, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Model. *The American Statistician* **54**(3), 217-224.

[40] MacKinnon, J. and White, H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics* **29**, 305–325.

[41] Mammen, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics* **21** (1), 255–285.

[42] Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models *The Annals of Statistics* **24** (1), 307-335.

[43] McCarthy, D. and Zhang, K. and Berk, R. and Brown, L. and Buja, A. and George, E. and Zhao, L. (2016). Calibrated Percentile Double Bootstrap For Robust Linear Regression Inference. *http://arxiv.org/abs/1511.00273*

[44] Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* **55** (3), 703-708.

[45] O'Hagan, A. (2013). Bayesian Inference with Misspecified Models: Inference about what? *Journal of Statistical Planning and Inference* **143**, 1643–1648.

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

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

[48] 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.

[49] Szpiro, A. A. and Rice, K. M. and Lumley, T. (2010). Model-Robust Regression and a Bayesian "Sandwich" Estimator. *The Annals of Applied Statistics* **4** (4), 2099-2113.

[50] Lee, J. D. and Sun, D. L. and Sun, Y. and Taylor, J. E. (2016). Exact Post-Selection Inference, with Application to the Lasso. *The Annals of Statistics* **44** (3), 907–927.

[51] Walker, S. G. (2013). Bayesian Inference with Misspecified Models. *Journal of Statistical Planning and Inference* **143**, 1621–1633.

[52] Wasserman, L. (2011). Low Assumptions, High Dimensions. *Rationality, Markets and Morals (RMM)* **2** (11), 201–209 (www.rmm-journal.de).

[53] Weber, N.C. (1986). The Jackknife and Heteroskedasticity (Consistent Variance Estimation for Regression Models). *Economics Letters* **20**, 161-163.

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

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

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

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

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

[59] Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics* **14** (4), 1261–1295.

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	t_{lin}	t_{boot}	t_{sand}
(Intercept)	36.459	5.103	8.038	8.145	**1.575**	**1.596**	1.013	7.144	4.536	4.477
CRIM	-0.108	0.033	0.035	0.031	1.055	0.945	0.896	-3.287	-3.115	-3.478
ZN	0.046	0.014	0.014	0.014	1.005	1.011	1.006	3.382	3.364	3.345
INDUS	0.021	0.061	0.051	0.051	**0.832**	**0.823**	0.990	0.334	0.402	0.406
CHAS	2.687	0.862	1.307	1.310	**1.517**	**1.521**	1.003	3.118	2.056	2.051
NOX	-17.767	3.820	3.834	3.827	1.004	1.002	0.998	-4.651	-4.634	-4.643
RM	3.810	0.418	0.848	0.861	**2.030**	**2.060**	1.015	9.116	4.490	4.426
AGE	0.001	0.013	0.016	0.017	1.238	1.263	1.020	0.052	0.042	0.042
DIS	-1.476	0.199	0.214	0.217	1.075	1.086	1.010	-7.398	-6.882	-6.812
RAD	0.306	0.066	0.063	0.062	0.949	0.940	0.990	4.613	4.858	4.908
TAX	-0.012	0.004	0.003	0.003	**0.736**	**0.723**	0.981	-3.280	-4.454	-4.540
PTRATIO	-0.953	0.131	0.118	0.118	0.899	0.904	1.005	-7.283	-8.104	-8.060
B	0.009	0.003	0.003	0.003	1.026	1.009	0.984	3.467	3.379	3.435
LSTAT	-0.525	0.051	0.100	0.101	**1.980**	**1.999**	1.010	-10.347	-5.227	-5.176

TABLE 6

Boston Housing data: Comparison of Standard Errors.

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	t_{lin}	t_{boot}	t_{sand}
(Intercept)	37.481	2.368	2.602	2.664	1.099	1.125	1.024	15.828	14.405	14.069
CRIM	4.179	1.746	1.539	1.533	0.882	0.878	0.996	2.394	2.715	2.726
ZN	0.826	1.418	1.359	1.353	0.959	0.954	0.995	0.583	0.608	0.611
INDUS	-1.844	1.501	1.410	1.413	0.939	0.941	1.002	-1.228	-1.308	-1.305
CHAS	6.328	1.764	2.490	2.485	**1.411**	**1.409**	0.998	3.587	2.542	2.547
NOX	-6.209	1.986	2.035	2.037	1.025	1.026	1.001	-3.127	-3.051	-3.048
RM	4.848	1.044	1.354	1.380	1.297	1.322	1.019	4.645	3.581	3.514
AGE	2.925	1.454	1.897	1.904	1.305	1.310	1.004	2.012	1.542	1.536
DIS	-9.047	1.754	1.933	1.945	1.102	1.109	1.006	-5.159	-4.679	-4.652
RAD	1.042	1.307	1.115	1.128	0.853	0.863	1.011	0.797	0.935	0.924
TAX	-5.319	1.343	1.155	1.157	0.860	0.862	1.003	-3.961	-4.607	-4.596
PTRATIO	-4.720	0.954	0.982	0.982	1.029	1.029	1.000	-4.946	-4.806	-4.808
B	-1.103	0.822	0.798	0.800	0.970	0.972	1.002	-1.342	-1.383	-1.380
LSTAT	-21.802	1.377	2.259	2.318	**1.641**	**1.683**	1.026	-15.832	-9.649	-9.404

TABLE 7

Boston Housing data: Comparison of Standard Errors; regressors are transformed with cdfs.

APPENDIX A: THE BOSTON HOUSING DATA

Table 6 illustrates discrepancies between types of standard errors with the Boston Housing data (Harrison and Rubinfeld 1978) which will be well known to many readers. Again, we dispense with the question as to whether the analysis is meaningful and focus on the comparison of standard errors. Here, too, $\boldsymbol{SE}_{boot}$ and $\boldsymbol{SE}_{sand}$ are mostly in agreement as they fall within less than 2% of each other, an exception being CRIM with a deviation of about 10%. By contrast, $\boldsymbol{SE}_{boot}$ and $\boldsymbol{SE}_{sand}$ are larger than their linear models cousin $\boldsymbol{SE}_{lin}$ by a factor of about 2 for RM and LSTAT, and about 1.5 for the intercept and the dummy variable CHAS. On the opposite side, $\boldsymbol{SE}_{boot}$ and $\boldsymbol{SE}_{sand}$ are less than 3/4 of $\boldsymbol{SE}_{lin}$ for TAX. For several regressors there is no major discrepancy among all three standard errors: ZN, NOX, B, and even for CRIM, $\boldsymbol{SE}_{lin}$ falls between the slightly discrepant values of $\boldsymbol{SE}_{boot}$ and $\boldsymbol{SE}_{sand}$.

Table 7 compares standard errors after the regressors are transformed to approximately uniform distributions using a rank or cdf transform.

Table 8 illustrates the $\boldsymbol{RAV}$ test for the Boston Housing data. Values of $\hat{\boldsymbol{RAV}}_j$

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{sand}$	$\boldsymbol{R\hat{A}V}_j$	2.5% Perm.	97.5% Perm.
(Intercept)	36.459	5.103	8.145	2.458*	0.859	1.535
CRIM	-0.108	0.033	0.031	0.776	0.511	3.757
ZN	0.046	0.014	0.014	1.006	0.820	1.680
INDUS	0.021	0.061	0.051	0.671*	0.805	1.957
CHAS	2.687	0.862	1.310	2.255*	0.722	1.905
NOX	-17.767	3.820	3.827	0.982	0.848	1.556
RM	3.810	0.418	0.861	4.087*	0.793	1.816
AGE	0.001	0.013	0.017	1.553*	0.860	1.470
DIS	-1.476	0.199	0.217	1.159	0.852	1.533
RAD	0.306	0.066	0.062	0.857	0.830	1.987
TAX	-0.012	0.004	0.003	0.512*	0.767	1.998
PTRATIO	-0.953	0.131	0.118	0.806*	0.872	1.402
B	0.009	0.003	0.003	0.995	0.786	1.762
LSTAT	-0.525	0.051	0.101	3.861*	0.803	1.798

TABLE 8

Boston Housing data: Permutation Inference for $\boldsymbol{R\hat{A}V}_j$ *(10,000 permutations).*

that fall outside the middle 95% range of their permutation null distributions are marked with asterisks.

APPENDIX B: ANCILLARITY

The facts as laid out in Section 4 amount to an argument against conditioning on regressors in regression. The justification for conditioning derives from an ancillarity argument according to which the regressors, if random, form an ancillary statistic for the linear model parameters $\boldsymbol{\beta}$ and σ^2, hence conditioning on $\mathbf{X}$ produces valid frequentist inference for these parameters (Cox and Hinkley 1974, Example 2.27). Indeed, with a suitably general definition of ancillarity, it can be shown that in *any* regression model the regressors form an ancillary. To see this we need an extended definition of ancillarity that includes nuisance parameters. The ingredients and conditions are as follows:

(1) $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$: the parameters, where $\boldsymbol{\psi}$ is of interest and $\boldsymbol{\lambda}$ is nuisance;
(2) $\boldsymbol{S} = (\boldsymbol{T}, \boldsymbol{A})$: a sufficient statistic with values $(\boldsymbol{t}, \boldsymbol{a})$;
(3) $p(\boldsymbol{t}, \boldsymbol{a};\, \boldsymbol{\psi}, \boldsymbol{\lambda}) = p(\boldsymbol{t} \,|\, \boldsymbol{a};\, \boldsymbol{\psi})\; p(\boldsymbol{a};\, \boldsymbol{\lambda})$: the condition that makes $\boldsymbol{A}$ an ancillary.

We say that the statistic $\boldsymbol{A}$ is ancillary for the parameter of interest, $\boldsymbol{\psi}$, in the presence of the nuisance parameter, $\boldsymbol{\lambda}$. Condition (3) can be interpreted as saying that the distribution of $\boldsymbol{T}$ is a mixture with mixing distribution $p(\boldsymbol{a}|\boldsymbol{\lambda})$. More importantly, for a fixed but unknown value $\boldsymbol{\lambda}$ and two values $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_0$, the likelihood ratio

$$\frac{p(\boldsymbol{t}, \boldsymbol{a};\, \boldsymbol{\psi}_1, \boldsymbol{\lambda})}{p(\boldsymbol{t}, \boldsymbol{a};\, \boldsymbol{\psi}_0, \boldsymbol{\lambda})} \;=\; \frac{p(\boldsymbol{t} \,|\, \boldsymbol{a};\, \boldsymbol{\psi}_1)}{p(\boldsymbol{t} \,|\, \boldsymbol{a};\, \boldsymbol{\psi}_0)}$$

has the nuisance parameter $\boldsymbol{\lambda}$ eliminated, justifying the conditionality principle according to which valid inference for $\boldsymbol{\psi}$ can be obtained by conditioning on $\boldsymbol{A}$.

When applied to regression, the principle implies that in *any* regression model the regressors, when random, are ancillary and hence can be conditioned on:

$$p(\boldsymbol{y}, \mathbf{X};\, \boldsymbol{\theta}) \;=\; p(\boldsymbol{y} \,|\, \mathbf{X};\, \boldsymbol{\theta})\; p_{\mathbf{X}}(\mathbf{X}),$$

where $\mathbf{X}$ acts as the ancillary $\boldsymbol{A}$ and $p_{\mathbf{X}}$ as the mixing distribution $p(\boldsymbol{a} \,|\boldsymbol{\lambda})$ with a "nonparametric" nuisance parameter that allows largely arbitrary distributions for the regressors. (The regressor distribution should grant identifiability of $\boldsymbol{\theta}$ in general, and non-collinearity in linear models in particular.) The literature does not seem to be rich in crisp definitions of ancillarity, but see, for example, Cox and Hinkley (1974, p.32-33). For the interesting history of ancillarity see the articles by Stigler (2001) and Aldrich (2005).

As explained in Section 4, the problem with the ancillarity argument is that it holds only when the regression model is correct. In practice, whether models are correct is never known.

APPENDIX C: ADJUSTMENT

C.1 Adjustment in Populations

To define the population-adjusted regressor random variable $X_{j\bullet}$, collect all other regressors in the random p-vector

$$\vec{\boldsymbol{X}}_{\text{-}j} = (1, X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)',$$

and let

$$X_{j\bullet} \;=\; X_j - \vec{\boldsymbol{X}}_{\text{-}j}{}'\,\boldsymbol{\beta}_{\text{-}j\bullet}, \qquad \text{where} \;\; \boldsymbol{\beta}_{\text{-}j\bullet} = \boldsymbol{E}[\,\vec{\boldsymbol{X}}_{\text{-}j}\vec{\boldsymbol{X}}_{\text{-}j}{}'\,]^{-1}\boldsymbol{E}[\,\vec{\boldsymbol{X}}_{\text{-}j}X_j].$$

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.

C.2 Adjustment in Samples

Define the sample-adjusted regressor column $\mathbf{X}_{j\hat{\bullet}}$ by collecting all regressor columns other than $\mathbf{X}_j$ in a $N \times p$ random regressor matrix

$$\mathbf{X}_{\text{-}j} = [\mathbf{1}, ..., \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, ..., \mathbf{X}_p]$$

and let

$$\mathbf{X}_{j\hat{\bullet}} \;=\; \mathbf{X}_j - \mathbf{X}_{\text{-}j}\,\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} \qquad \text{where} \;\; \hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} = (\mathbf{X}_{\text{-}j}{}'\,\mathbf{X}_{\text{-}j})^{-1}\mathbf{X}_{\text{-}j}{}'\,\mathbf{X}_j.$$

(Note the use of hat notation "$\hat{\bullet}$" to distinguish it from population-based adjustment "$\bullet$".) 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.

APPENDIX D: PROOFS

D.1 Precise Non-Ancillarity Statements and Proofs for Section 4

Lemma: *The functional $\boldsymbol{\beta}(\boldsymbol{P})$ depends on $\boldsymbol{P}$ only through the conditional mean function and the regressor distribution; it does not depend on the conditional noise distribution.*

In the nonlinear case the clause $\exists \boldsymbol{P}_1, \boldsymbol{P}_2 : \boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$ is driven solely by differences in the regressor distributions $\boldsymbol{P}_1(\mathrm{d}\vec{\boldsymbol{x}})$ and $\boldsymbol{P}_2(\mathrm{d}\vec{\boldsymbol{x}})$ because $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ share the mean function $\mu_0(.)$ while their conditional noise distributions are irrelevant by the above lemma.

The Lemma is more precisely stated as follows: For two data distributions $\boldsymbol{P}_1(\mathrm{d}y, \mathrm{d}\vec{\boldsymbol{x}})$ and $\boldsymbol{P}_2(\mathrm{d}y, \mathrm{d}\vec{\boldsymbol{x}})$ the following holds:

$$\boldsymbol{P}_1(\mathrm{d}\vec{\boldsymbol{x}}) = \boldsymbol{P}_2(\mathrm{d}\vec{\boldsymbol{x}}), \quad \mu_1(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}_{1,2}}{=} \mu_2(\vec{\boldsymbol{X}}) \quad \Longrightarrow \quad \boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}(\boldsymbol{P}_2).$$

Proposition: *The OLS functional $\boldsymbol{\beta}(\boldsymbol{P})$ does* **not** *depend on the regressor distribution if and only if $\mu(\vec{\boldsymbol{X}})$ is linear. More precisely, for a fixed measurable function $\mu_0(\vec{\boldsymbol{x}})$ consider the class of data distributions $\boldsymbol{P}$ for which $\mu_0(.)$ is a version of their conditional mean function: $\boldsymbol{E}[Y|\vec{\boldsymbol{X}}] = \mu(\vec{\boldsymbol{X}}) \stackrel{\boldsymbol{P}}{=} \mu_o(\vec{\boldsymbol{X}})$. In this class the following holds:*

$$\begin{array}{lcl} \mu_0(.) \text{ is nonlinear} & \Longrightarrow & \exists \boldsymbol{P}_1, \boldsymbol{P}_2 : \;\; \boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2), \\ \mu_0(.) \text{ is linear} & \Longrightarrow & \forall \boldsymbol{P}_1, \boldsymbol{P}_2 : \;\; \boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}(\boldsymbol{P}_2). \end{array}$$

For the proposition we show the following: For a fixed measurable function $\mu_0(\vec{\boldsymbol{x}})$ consider the class of data distributions $\boldsymbol{P}$ for which $\mu_0(.)$ is a version of their conditional mean function: $\boldsymbol{E}[Y|\vec{\boldsymbol{X}}] = \mu(\vec{\boldsymbol{X}}) \stackrel{\boldsymbol{P}}{=} \mu_o(\vec{\boldsymbol{X}})$. In this class the following holds:

$$\begin{array}{lcl} \mu_0(.) \text{ is nonlinear} & \Longrightarrow & \exists \boldsymbol{P}_1, \boldsymbol{P}_2 : \;\; \boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2), \\ \mu_0(.) \text{ is linear} & \Longrightarrow & \forall \boldsymbol{P}_1, \boldsymbol{P}_2 : \;\; \boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}(\boldsymbol{P}_2). \end{array}$$

The linear case is trivial: if $\mu_0(\vec{\boldsymbol{X}})$ is linear, that is, $\mu_0(\vec{\boldsymbol{x}}) = \boldsymbol{\beta}' \vec{\boldsymbol{x}}$ for some $\boldsymbol{\beta}$, then $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}$ irrespective of $\boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}})$. The nonlinear case is proved as follows: For any set of points $\vec{\boldsymbol{x}}_1, ...\vec{\boldsymbol{x}}_{p+1} \in \mathbb{R}^{p+1}$ in general position and with 1 in the first coordinate, there exists a unique linear function $\boldsymbol{\beta}' \vec{\boldsymbol{x}}$ through the values of $\mu_0(\vec{\boldsymbol{x}}_i)$. Define $\boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}})$ by putting mass $1/(p+1)$ on each point; define the conditional distribution $\boldsymbol{P}(\mathrm{d}y \,|\, \vec{\boldsymbol{x}}_i)$ as a point mass at $y = \mu_o(\vec{\boldsymbol{x}}_i)$; this defines $\boldsymbol{P}$ such that $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}$. Now, if $\mu_0()$ is nonlinear, there exist two such sets of points with differing linear functions $\boldsymbol{\beta}_1{}' \vec{\boldsymbol{x}}$ and $\boldsymbol{\beta}_2{}' \vec{\boldsymbol{x}}$ to match the values of $\mu_0()$ on these two sets; by following the preceding construction we obtain $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ such that $\boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 = \boldsymbol{\beta}(\boldsymbol{P}_2)$.

D.2 *RAV* Decomposition

Lemma D.2: ***RAV*** *Decomposition.*

$$\boldsymbol{RAV}[\hat{\beta}_j, m^2] \;=\; w_\sigma \, \boldsymbol{RAV}[\hat{\beta}_j, \sigma^2] \;+\; w_\eta \, \boldsymbol{RAV}[\hat{\beta}_j, \eta^2],$$

$$\text{where} \quad w_\sigma := \frac{E[\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}, \quad w_\eta := \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}, \quad w_\sigma + w_\eta = 1.$$

D.3 Proof of the *RAV*-Range Proposition in Section 11.5

Proposition D.3: *If $\boldsymbol{E}[X_{j\bullet}^2] < \infty$, then*

$$\sup_{m_j^2} \boldsymbol{RAV}[\hat{\beta}_j, m_j^2] \;=\; \frac{\boldsymbol{P}\text{-}\max X_{j\bullet}^2}{\boldsymbol{E}[X_{j\bullet}^2]}, \qquad \inf_{m_j^2} \boldsymbol{RAV}[\hat{\beta}_j, m_j^2] \;=\; \frac{\boldsymbol{P}\text{-}\min X_{j\bullet}^2}{\boldsymbol{E}[X_{j\bullet}^2]}.$$

Here are some corollaries that follow from the proposition:

- If, for example, $X_{j\bullet} \sim U[-1, +1]$ is uniformly distributed, then $\boldsymbol{E}[X_{j\bullet}^2] = 1/3$. Hence the upper bound on the ***RAV*** is 3 and, asymptotically, the usual standard error will never be too short by more than a factor $\sqrt{3} \approx 1.732$.

- However, when $\boldsymbol{E}[X_{j\bullet}^2]$ is very small compared to $\boldsymbol{P}\text{-max}\, X_{j\bullet}^2$, that is, when $X_{j\bullet}$ is highly concentrated around its mean 0, then this approximates the case of an unbounded support and the worst-case $\boldsymbol{RAV}$ can be very large.
- If, on the other hand, $\boldsymbol{E}[X_{j\bullet}^2]$ is very close to $\boldsymbol{P}\text{-max}\, X_{j\bullet}^2 = c^2$, then $X_{j\bullet}$ approximates a balanced two-point distribution at $\pm c$, and the sandwich and usual standard errors necessarily agree in the limit.

The result for the last case, a two-point balanced distribution, is intuitive because here it is impossible to detect nonlinearity. Heteroskedasticity, however, is still possible (different noise variances at $\pm c$), but this does not matter because the dependence of $\boldsymbol{RAV}$ is on $X_{j\bullet}^2$, not $X_{j\bullet}$, and $X_{j\bullet}^2$ has a one-point distribution at c^2. The $\boldsymbol{RAV}$ can only respond to heteroskedasticities that vary in $X_{j\bullet}^2$.

The $\boldsymbol{RAV}$ is a functional of $X_{j\bullet}^2$ and $f_j^2(X_{j\bullet}^2)$, suggesting simplified notation: X^2 for $X_{j\bullet}^2$, $f^2(X^2)$ for $f_j^2(X_{j\bullet}^2)$, and $\boldsymbol{RAV}[f^2]$ for $\boldsymbol{RAV}[\hat{\beta}_j, f_j^2]$. Proposition D.3 is proved by the first lemma as applied to $\sigma_j^2(X_{j\bullet}^2)$, and by the second lemma as applied to $\eta_j^2(X_{j\bullet}^2)$. The difference between the two cases is that nonlinearities $\eta_j(X_{j\bullet}^2)$ is necessarily centered whereas for $\sigma_j^2(X_{j\bullet}^2)$ there exists no such requirement; the construction below requires in the centered case that $\boldsymbol{P}\text{-min}$ and $\boldsymbol{P}\text{-max}$ of $X_{j\bullet}^2$ do not carry positive probability mass. This is a largely technical condition because even for discrete regressors X_j the adjusted squared version $X_{j\bullet}^2$ will have a continuous distribution if there exists just one other regressor that is continuous and non-orthogonal (partly collinear) to X_j.

Lemma D.3.1: *Assume* $\boldsymbol{E}[X^2] < \infty$.

(a) Define a one-parameter family f_t^2*:*

$$f_t^2(X^2) \;:=\; \frac{1_{[|X| \geq t]}}{p(t)}\,, \qquad \textit{where} \quad p(t) := \boldsymbol{P}[|X| \geq t]$$

for $p(t) > 0$. *Then the following holds:*

$$\sup_t \boldsymbol{RAV}[f_t^2] \;=\; \frac{\boldsymbol{P}\text{-max}\, X^2}{\boldsymbol{E}[X^2]}.$$

(b) Define a one-parameter family g_t^2*:*

$$g_t^2(X^2) \;:=\; \frac{1_{[|X| \leq t]}}{\bar{p}(t)}\,, \qquad \textit{where} \quad \bar{p}(t) := \boldsymbol{P}[|X| \leq t]\,.$$

Then the following holds:

$$\inf_t \boldsymbol{RAV}[g_t^2] \;=\; \frac{\boldsymbol{P}\text{-min}\, X^2}{\boldsymbol{E}[X^2]}.$$

Proof of part (a): Preliminary observations:

- $\boldsymbol{E}[f_t^2(X^2)] = 1$.
- $\boldsymbol{E}[f_t^2(X^2) X^2] \leq \boldsymbol{P}\text{-max}\, X^2$.
- $\boldsymbol{P}\text{-max}\, X^2 = \sup_{p(t)>0} t^2$.

For $p(t) > 0$ we have

$$\boldsymbol{E}\left[f_t^2(X)X^2\right] \;=\; \frac{1}{p(t)}\,\boldsymbol{E}\left[1_{[|X|\geq t]}\,X^2\right] \;\geq\; \frac{1}{p(t)}\,p(t)\,t^2 \;=\; t^2,$$

hence $\sup_t \boldsymbol{E}\left[f_t^2(X)X^2\right] = \boldsymbol{P}\text{-}\max X^2$. $\square$

Proof of part (b): Preliminary observations:

- $\boldsymbol{E}[g_t^2(X^2)] = 1$.
- $\boldsymbol{E}[g_t^2(X^2)X^2] \geq \boldsymbol{P}\text{-}\min X^2$.
- $\boldsymbol{P}\text{-}\min X^2 = \inf_{\bar{p}(t)>0} t^2$.

For $\bar{p}(t) > 0$ we have:

$$\boldsymbol{E}\left[g_t^2(X)X^2\right] \;=\; \frac{1}{\bar{p}(t)}\,\boldsymbol{E}\left[1_{[|X|\leq t]}\,X^2\right] \;\leq\; \frac{1}{\bar{p}(t)}\,\bar{p}(t)\,t^2 \;=\; t^2,$$

hence $\inf_t \boldsymbol{E}\left[g_t^2(X)X^2\right] = \boldsymbol{P}\text{-}\min X^2$. $\square$

Lemma D.3.2:

(a) Define a one-parameter family

$$f_t(X^2) \;=\; \frac{1_{[|X|\geq t]} - p(t)}{\sqrt{p(t)(1-p(t))}}\,, \qquad \textit{where} \quad p(t) = \boldsymbol{P}[|X| \geq t]\,,$$

for $p(t)>0$ and $1-p(t)>0$. If $p(t)$ is continuous at $t=\boldsymbol{P}\text{-}\max|X|$, that is, $\boldsymbol{P}[|X| = \boldsymbol{P}\text{-}\max|X|] = 0$, then

$$\sup_t \boldsymbol{RAV}[f_t^2] \;=\; \frac{\boldsymbol{P}\text{-}\max X^2}{\boldsymbol{E}[X^2]}.$$

(b) Define a one-parameter family

$$g_t(X^2) \;=\; \frac{1_{[|X|\leq t]} - \bar{p}(t)}{\sqrt{\bar{p}(t)(1-\bar{p}(t))}}\,, \qquad \textit{where} \quad \bar{p}(t) = \boldsymbol{P}[|X| \leq t]\,,$$

for $\bar{p}(t)>0$ and $1-\bar{p}(t)>0$. If $\bar{p}(t)$ is continuous at $t=\boldsymbol{P}\text{-}\min|X|$, that is, $\boldsymbol{P}[|X| = \boldsymbol{P}\text{-}\min|X|] = 0$, then

$$\inf_t \boldsymbol{RAV}[g_t^2] \;=\; \frac{\boldsymbol{P}\text{-}\min X^2}{\boldsymbol{E}[X^2]}.$$

Proof of part (a): Preliminary observations:

- $\boldsymbol{E}[f_t^2(X^2)] = 1$.
- $\boldsymbol{E}[f_t^2(X^2)X^2] \leq \boldsymbol{P}\text{-}\max X^2$.
- $\boldsymbol{P}\text{-}\max X^2 = \sup_{0<p(t)<1} t^2$.

For $p(t)>0$ we have:

$$\begin{aligned}
\boldsymbol{E}\left[f_t^2(X)X^2\right] &= \frac{1}{p(t)(1-p(t))}\,\boldsymbol{E}\left[\left(1_{[|X|\geq t]}-p(t)\right)^2 X^2\right] \\
&= \frac{1}{p(t)(1-p(t))}\,\left(\boldsymbol{E}\left[1_{[|X|\geq t]}X^2\right](1-2\,p(t))+p(t)^2\boldsymbol{E}[X^2]\right) \\
&\geq \frac{1}{p(t)(1-p(t))}\,\left(p(t)\,t^2\,(1-2\,p(t))\,+p(t)^2\boldsymbol{E}[X^2]\right) \qquad \text{for} \quad p(t)\leq\frac{1}{2} \\
&= \frac{1}{1-p(t)}\,\left(t^2\,(1-2\,p(t))+p(t)\boldsymbol{E}[X^2]\,\right) \\
&\longrightarrow \boldsymbol{P}\text{-}\max X^2
\end{aligned}$$

as $t\uparrow \boldsymbol{P}\text{-}\max|X|$ and hence $p(t)\downarrow 0$. $\square$

Proof of part (b): Preliminary observations:

- $\boldsymbol{E}[g_t^2(X^2)]=1$.
- $\boldsymbol{E}[g_t^2(X^2)X^2]\geq \boldsymbol{P}\text{-}\min X^2$.
- $\boldsymbol{P}\text{-}\min X^2=\inf_{0<\bar{p}(t)<1} t^2$.

$$\begin{aligned}
\boldsymbol{E}\left[g_t^2(X)^2X^2\right] &= \frac{1}{\bar{p}(t)(1-\bar{p}(t))}\,\boldsymbol{E}\left[\left(1_{[|X|\leq t]}-\bar{p}(t)\right)^2 X^2\right] \\
&= \frac{1}{\bar{p}(t)(1-\bar{p}(t))}\,\left(\boldsymbol{E}\left[1_{[|X|\leq t]}X^2(1-2\,\bar{p}(t))\right]+\bar{p}(t)^2\boldsymbol{E}[X^2]\right) \\
&\leq \frac{1}{\bar{p}(t)(1-\bar{p}(t))}\,\left(\bar{p}(t)\,t^2\,(1-2\,\bar{p}(t))+\bar{p}(t)^2\boldsymbol{E}[X^2]\right) \qquad \text{for} \quad \bar{p}(t)\leq\frac{1}{2} \\
&= \frac{1}{1-\bar{p}(t)}\,\left(t^2\,(1-2\,\bar{p}(t))+\bar{p}(t)\boldsymbol{E}[X^2]\,\right) \\
&\longrightarrow \boldsymbol{P}\text{-}\min X^2
\end{aligned}$$

as $t\downarrow \boldsymbol{P}\text{-}\min|X|$ and hence $\bar{p}(t)\downarrow 0$. $\square$

D.4 Details for Figure 6

We write X instead of $X_{j\bullet}$ and assume it has a standard normal distribution, $X\sim N(0,1)$, whose density will be denoted by $\phi(x)$. In Figure 6 the base function is, up to scale, as follows:

$$f(x) \;=\; \exp\left(-\frac{t}{2}\,\frac{\boldsymbol{x}^2}{2}\right), \qquad t>-1.$$

These functions are normal densities up to normalization for $t>0$, constant 1 for $t=0$, and convex for $t<0$. Conveniently, $f(x)\phi(x)$ and $f^2(x)\phi(x)$ are both normal densities (up to normalization) for $t>-1$:

$$\begin{aligned}
f(x)\,\phi(x) &= s_1\,\phi_{s_1}(x), \qquad & s_1=(1+t/2)^{-1/2}, \\
f^2(x)\,\phi(x) &= s_2\,\phi_{s_2}(x), \qquad & s_2=(1+t)^{-1/2},
\end{aligned}$$

where we write $\phi_s(x) = \phi(x/s)/s$ for scaled normal densities. Accordingly we obtain the following moments:

$$\begin{array}{lclclcl}
\boldsymbol{E}[f(X)] & = & s_1\,\boldsymbol{E}[\,1\,|N(0,{s_1}^2)] & = & s_1 & = & (1+t/2)^{-1/2}, \\
\boldsymbol{E}[f(X)\,X^2] & = & s_1\,\boldsymbol{E}[X^2|N(0,{s_1}^2)] & = & {s_1}^3 & = & (1+t/2)^{-3/2}, \\
\boldsymbol{E}[f^2(X)] & = & s_2\,\boldsymbol{E}[\,1\,|N(0,{s_2}^2)] & = & s_2 & = & (1+t)^{-1/2}, \\
\boldsymbol{E}[f^2(X)\,X^2] & = & s_2\,\boldsymbol{E}[X^2|N(0,{s_2}^2)] & = & {s_2}^3 & = & (1+t)^{-3/2},
\end{array}$$

and hence

$$\boldsymbol{RAV}[\hat{\beta}, f^2] \;=\; \frac{\boldsymbol{E}[f^2(X)\,X^2]}{\boldsymbol{E}[f^2(X)]\,\boldsymbol{E}[X^2]} \;=\; {s_2}^2 \;=\; (1+t)^{-1}$$

Figure 6 shows the functions as follows: $f(x)^2/\boldsymbol{E}[f^2(X)] = f(x)^2/s_2$.

D.5 Proof of Asymptotic Normality of $\hat{\boldsymbol{RAV}}_j$, Section 12.2

We will need notation for each observation's population-adjusted regressors: $\mathbf{X}_{j\bullet} = (X_{1,j\bullet}, ..., X_{N,j\bullet})' = \mathbf{X}_j - \mathbf{X}_{\text{-}j}\boldsymbol{\beta}_{\text{-}j\bullet}$. The following distinction is elementary but important: The component variables of $\mathbf{X}_{j\bullet} = (X_{i,j\bullet})_{i=1...N}$ are iid as they are population-adjusted, whereas the component variables of $\mathbf{X}_{j\hat{\bullet}} = (X_{i,j\hat{\bullet}})_{i=1...N}$ are dependent as they are sample-adjusted. As $N \to \infty$ for fixed p, this dependency disappears asymptotically, and we have for the empirical distribution of the values $\{X_{i,j\hat{\bullet}}\}_{i=1...N}$ the obvious convergence in distribution:

$$\{X_{i,j\hat{\bullet}}\}_{i=1...N} \quad \xrightarrow{\mathcal{D}} \quad X_{j\bullet} \;\overset{\mathcal{D}}{=}\; X_{i,j\bullet} \qquad (N \to \infty).$$

We recall (27) for reference in the following form:

$$\hat{\boldsymbol{RAV}}_j \;=\; \frac{\frac{1}{N}\langle(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^2, {\mathbf{X}_{j\hat{\bullet}}}^2\rangle}{\frac{1}{N}\|\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2\;\frac{1}{N}\|\mathbf{X}_{j\hat{\bullet}}\|^2}\,. \tag{29}$$

For the denominators it is easy to show that

$$\begin{array}{rcl}
\frac{1}{N}\|\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 & \xrightarrow{\boldsymbol{P}} & \boldsymbol{E}[\,\delta^2\,], \\
\frac{1}{N}\|\mathbf{X}_{j\hat{\bullet}}\|^2 & \xrightarrow{\boldsymbol{P}} & \boldsymbol{E}[\,X_{j\bullet}^2\,].
\end{array} \tag{30}$$

For the numerator a CLT holds based on

$$\frac{1}{N^{1/2}}\langle(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^2, {\mathbf{X}_{j\hat{\bullet}}}^2\rangle \;=\; \frac{1}{N^{1/2}}\langle(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^2, {\mathbf{X}_{j\bullet}}^2\rangle + O_P(N^{-1/2}). \tag{31}$$

For a proof outline see **Details** below. It is therefore sufficient to show asymptotic normality of $\langle\boldsymbol{\delta}^2, {\mathbf{X}_{j\bullet}}^2\rangle$. Here are first and second moments:

$$\begin{array}{rclcl}
\boldsymbol{E}[\frac{1}{N}\langle\boldsymbol{\delta}^2, {\mathbf{X}_{j\bullet}}^2\rangle] & = & \boldsymbol{E}[\delta^2\,X_{j\bullet}^2] & = & \boldsymbol{E}[\delta^2]\,\boldsymbol{E}[X_{j\bullet}^2], \\
\boldsymbol{V}[\frac{1}{N^{1/2}}\langle\boldsymbol{\delta}^2, {\mathbf{X}_{j\bullet}}^2\rangle] & = & \boldsymbol{E}[\delta^4\,{X_{j\bullet}}^4] - \boldsymbol{E}[\delta^2\,X_{j\bullet}^2]^2 & = & \boldsymbol{E}[\delta^4]\,\boldsymbol{E}[{X_{j\bullet}}^4] - \boldsymbol{E}[\delta^2]^2\,\boldsymbol{E}[X_{j\bullet}^2]^2.
\end{array}$$

The second equality on each line holds under the null hypothesis of independent δ and $\vec{\boldsymbol{X}}$. For the variance one observes that we assume that $\{(Y_i, \vec{\boldsymbol{X}}_i)\}_{i=1...N}$ to be iid sampled pairs, hence $\{(\delta_i^2, {X_{i,j\bullet}}^2)\}_{i=1...N}$ are N iid sampled pairs as well.

Using the denominator terms (30) and Slutsky's theorem, we arrive at the first version of the CLT for $\boldsymbol{R\hat{A}V}_j$:

$$N^{1/2}\,(\boldsymbol{R\hat{A}V}_j - 1) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(0,\, \frac{\boldsymbol{E}[\,\delta^4]}{\boldsymbol{E}[\,\delta^2]^2}\, \frac{\boldsymbol{E}[X_{j\bullet}{}^4]}{\boldsymbol{E}[X_{j\bullet}^2]^2} \,-\, 1\right)$$

With the additional null assumption of normal noise we have $\boldsymbol{E}[\,\delta^4] = 3\boldsymbol{E}[\,\delta^2]^2$, and hence the second version of the CLT for $\boldsymbol{R\hat{A}V}_j$:

$$N^{1/2}\,(\boldsymbol{R\hat{A}V}_j - 1) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(0,\, 3\,\frac{\boldsymbol{E}[X_{j\bullet}{}^4]}{\boldsymbol{E}[X_{j\bullet}^2]^2} - 1\right).$$

Details for the numerator (31), using notation of Sections C.1 and C.2, in particular $\mathbf{X}_{j\bullet} = \mathbf{X}_j - \mathbf{X}_{\text{-}j}\boldsymbol{\beta}_{\text{-}j\bullet}$ and $\mathbf{X}_{j\hat{\bullet}} = \mathbf{X}_j - \mathbf{X}_{\text{-}j}\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}}$:

(32)

$$\begin{aligned}
\langle (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^2, \mathbf{X}_{j\hat{\bullet}}{}^2 \rangle \;&=\; \langle\, ((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) - \mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2,\, (\mathbf{X}_{j\bullet} - \mathbf{X}_{\text{-}j}(\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \boldsymbol{\beta}_{\text{-}j\bullet}))^2\,\rangle \\
&=\; \langle\, \boldsymbol{\delta}^2 + (\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2 - 2\,\boldsymbol{\delta}\,(\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})), \\
&\qquad \mathbf{X}_{j\bullet}{}^2 + (\mathbf{X}_{\text{-}j}(\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \hat{\boldsymbol{\beta}}_{\text{-}j\bullet}))^2 - 2\,\mathbf{X}_{j\bullet}(\mathbf{X}_{\text{-}j}(\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \boldsymbol{\beta}_{\text{-}j\bullet}))\,\rangle \\
&=\; \langle\, \boldsymbol{\delta}^2, \mathbf{X}_{j\bullet}{}^2\,\rangle + \ldots
\end{aligned}$$

Among the 8 terms in "...", each contains at least one subterm of the form $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ or $\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \boldsymbol{\beta}_{\text{-}j\bullet}$, each being of order $O_P(N^{-1/2})$. We first treat the terms with just one of these subterms to first power, of which there are only two, normalized by $N^{1/2}$:

$$\begin{aligned}
\tfrac{1}{N^{1/2}}\,\langle\, -2\,\boldsymbol{\delta}\,(\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})),\, \mathbf{X}_{j\bullet}{}^2\,\rangle \;&=\; -2\textstyle\sum_{k=0\ldots p}\left(\tfrac{1}{N^{1/2}}\sum_{i=1\ldots N}\delta_i X_{i,k} X_{i,j\bullet}^2\right)(\hat{\beta}_j - \beta_j) \\
&=\; \textstyle\sum_{k=0\ldots p} O_P(1)\,O_P(N^{-1/2}) \;=\; O_P(N^{-1/2}), \\
\tfrac{1}{N^{1/2}}\,\langle\, \boldsymbol{\delta}^2,\, -2\,\mathbf{X}_{j\bullet}(\mathbf{X}_{\text{-}j}(\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \boldsymbol{\beta}_{\text{-}j\bullet}))\,\rangle \;&=\; -2\textstyle\sum_{k(\neq j)}\left(\tfrac{1}{N^{1/2}}\sum_{i=1\ldots N}\delta_i^2 X_{i,j\bullet} X_{i,k}\right)(\hat{\beta}_{\text{-}j\hat{\bullet},k} - \beta_{\text{-}j\bullet,k}) \\
&=\; \textstyle\sum_{k(\neq j)} O_P(1)\,O_P(N^{-1/2}) \;=\; O_P(N^{-1/2}).
\end{aligned}$$

The terms in the big parens are $O_P(1)$ because they are asymptotically normal. This is so because they are centered under the null hypothesis that δ_i is independent of the regressors $\vec{\boldsymbol{X}}_i$: In the first term we have

$$\boldsymbol{E}[\delta_i X_{i,k} X_{i,j\bullet}^2] = \boldsymbol{E}[\delta_i]\,\boldsymbol{E}[X_{i,k} X_{i,j\bullet}^2] = 0$$

due to $\boldsymbol{E}[\delta_i] = 0$. In the second term we have

$$\boldsymbol{E}[\delta_i^2 X_{i,j\bullet} X_{i,k}] = \boldsymbol{E}[\delta_i^2]\,\boldsymbol{E}[X_{i,j\bullet} X_{i,k}] = 0$$

due to $\boldsymbol{E}[X_{i,j\bullet} X_{i,k}] = 0$ as $k \neq j$.

We proceed to the 6 terms in (32) that contain at least two β-subterms or one β-subterm squared. For brevity we treat one term in detail and assume that the reader will be convinced that the other 5 terms can be dealt with similarly. Here is one such term, again scaled for CLT purposes:

$$\begin{aligned}
\tfrac{1}{N^{1/2}}\langle\, (\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2, \mathbf{X}_{j\bullet}{}^2\,\rangle \;&=\; \textstyle\sum_{k,l=0\ldots p}\left(\tfrac{1}{N}\sum_{i=1\ldots N} X_{i,k} X_{i,l} X_{i,j\bullet}^2\right) N^{1/2}(\hat{\beta}_k - \beta_k)(\hat{\beta}_l - \beta_l) \\
&=\; \textstyle\sum_{k,l=0\ldots p} \mathrm{const}\cdot O_P(1)\,O_P(N^{-1/2}) \;=\; O_P(N^{-1/2}).
\end{aligned}$$

The term in the parens converges in probability to $\boldsymbol{E}[X_{i,k}X_{i,l}X_{i,j\bullet}^2]$, accounting for "const"; the term $N^{1/2}(\hat{\beta}_k - \beta_k)$ is asymptotically normal and hence $O_P(1)$; and the term $(\hat{\beta}_l - \beta_l)$ is $O_P(N^{-1/2})$ due to its CLT.

Details for the denominator terms (30): It is sufficient to consider the first denominator term. Let $\boldsymbol{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the hat or projection matrix for $\mathbf{X}$.

$$\begin{aligned}
\tfrac{1}{N}\,\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 &= \tfrac{1}{N}\,\mathbf{Y}'\,(\boldsymbol{I} - \boldsymbol{H})\mathbf{Y} \\
&= \tfrac{1}{N}\,\left(\|\mathbf{Y}\|^2 - \mathbf{Y}'\,\boldsymbol{H}\mathbf{Y}\right) \\
&= \tfrac{1}{N}\,\|\mathbf{Y}\|^2 \;-\; \left(\tfrac{1}{N}\sum Y_i\vec{\boldsymbol{X}}_i{}'\right)\left(\tfrac{1}{N}\sum \vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i{}'\right)^{-1}\left(\tfrac{1}{N}\sum \vec{\boldsymbol{X}}_i Y_i\right) \\
&\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[Y^2] \;-\; \boldsymbol{E}[Y\vec{\boldsymbol{X}}]\,\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}Y] \\
&= \boldsymbol{E}[Y^2] - \boldsymbol{E}[Y\vec{\boldsymbol{X}}'\,\boldsymbol{\beta}] \\
&= \boldsymbol{E}[(Y - \vec{\boldsymbol{X}}'\,\boldsymbol{\beta})^2] \qquad \text{due to } \;\boldsymbol{E}[(Y - \vec{\boldsymbol{X}}'\,\boldsymbol{\beta})\vec{\boldsymbol{X}}] = \mathbf{0} \\
&= \boldsymbol{E}[\,\delta^2].
\end{aligned}$$

The calculations are the same for the second denominator term, substituting $\mathbf{X}_j$ for $\mathbf{Y}$, $\mathbf{X}_{\text{-}j}$ for $\mathbf{X}$, $X_{j\bullet}$ for δ, and $\boldsymbol{\beta}_{\text{-}j\bullet}$ for $\boldsymbol{\beta}$.

APPENDIX E: NON-NORMALITY OF CONDITIONAL NULL DISTRIBUTIONS OF $\hat{RAV}_J$

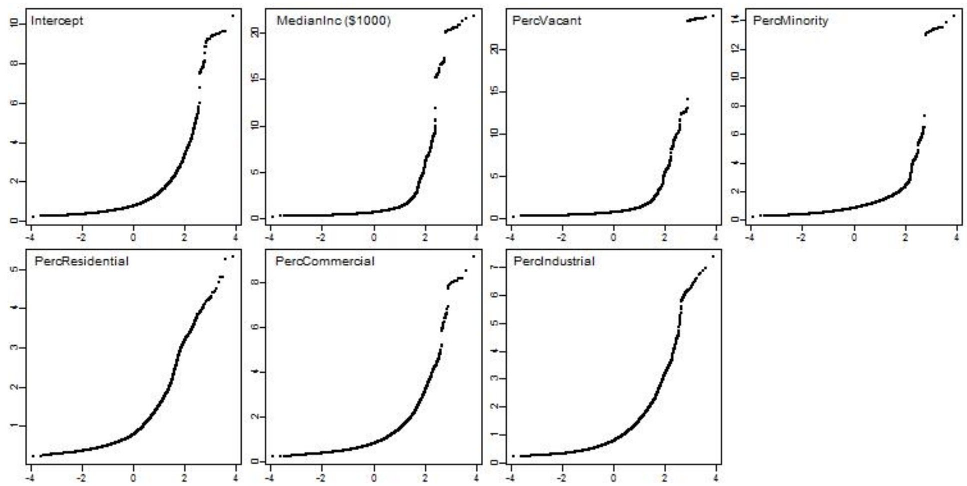

FIG 9. *Permutation distributions of* $\boldsymbol{R\hat{A}V}_j$ *for the LA Homeless Data*

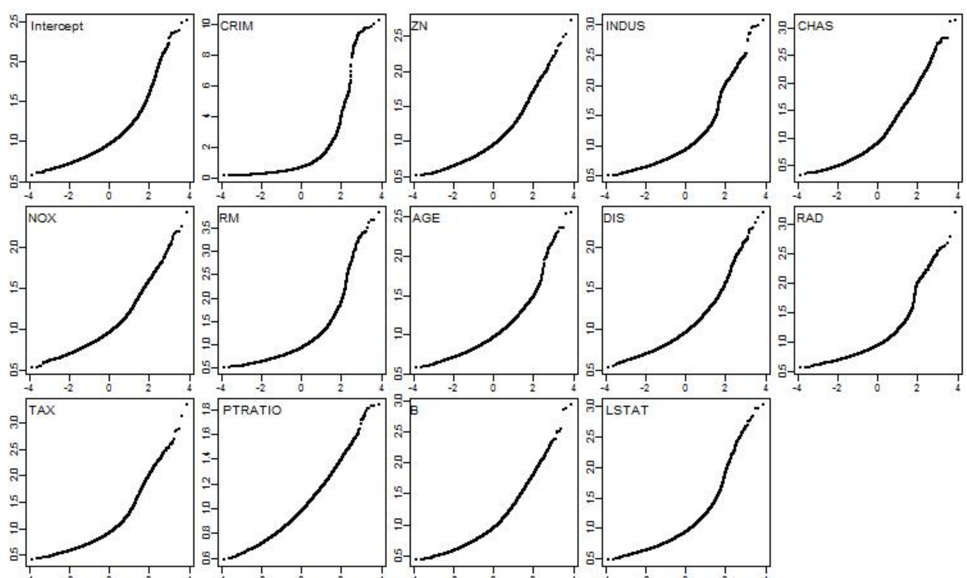

FIG 10. *Permutation distributions of* $\boldsymbol{R\hat{A}V}_j$ *for the Boston Housing Data*