Submitted to Statistical Science

Models as Approximations — A Conspiracy of Random Regressors and Model Deviations Against Classical Inference in Regression

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

Wharton – University of Pennsylvania[‡] and Amazon.com[§] 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. It is asymptotically correct even under "model misspecification," that is, when models are approximations rather than generative truths. It is well-known to be "heteroskedasticity-consistent", but it is less well-known to be "nonlinearity-consistent" as well. Nonlinearity, however, raises fundamental issues: When fitted models are approximations, conditioning on the regressor is no longer permitted because the ancillarity argument that justifies it breaks down. Two effects occur: (1) parameters become dependent on the regressor distribution; (2) the sampling variability of parameter estimates no longer derives from the conditional distribution of the response alone. Additional sampling variability arises when the nonlinearity conspires with the randomness of the regressors to generate a $1/\sqrt{N}$ contribution to standard errors. Asymptotically, standard errors from "model-trusting" fixed-regressor theories can deviate from those of "model-robust" random-regressor theories by arbitrary magnitudes. In the case of linear models, a test will be proposed for comparing the two types of standard errors.

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). – Amazon.com. – 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 given by a regressor matrix $\boldsymbol{X}_{N\times(p+1)}$ and a 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, $\boldsymbol{V}[\boldsymbol{y}]=\boldsymbol{D}$ diagonal:

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

The right hand side has the characteristic "sandwich" form, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ forming the "bread" and $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}$ 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:

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

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:

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

The term "heteroskedasticity-consistent" is an unfortunate choice as it obscures the fact that the same estimator of standard error is also "nonlinearity-consistent." Because of the relative obscurity of this important fact we will pay considerable attention to its implications. In particular we 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; see Figures 2 and 4 below. A more striking illustration is available to users of the **R** *Language* by executing the following line of code:

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

Side remarks:

- The term "nonlinearity" is meant in the sense of (3), first order model deviation, $\boldsymbol{E}[\,\boldsymbol{y}\,|\boldsymbol{X}]-\boldsymbol{X}\boldsymbol{\beta}\neq\boldsymbol{0}$. A different meaning of "nonlinearity", *not* intended here, occurs when the regressor matrix $\boldsymbol{X}$ contains multiple columns that are functions (polynomials, B-splines, ...) of an independent variable. We distinguish between "regressors" and "independent variables": Multiple regressors may be functions of the same independent variable.
- 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). Also, generalizations are pervasive in Generalized Estimating Equations (GEE; Liang and Zeger 1986; Diggle et al. 2002) and Generalized Method of Moments (GMM; Hansen 1982).

From the sandwich estimator (2), the "usual" estimator of linear models theory is obtained by collapsing the sandwich form assuming homoskedasticity:

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

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: $\boldsymbol{E}[\,\boldsymbol{y}\,|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$ (linearity) and $\boldsymbol{V}[\,\boldsymbol{y}\,|\boldsymbol{X}] = \sigma^2\boldsymbol{I}_N$ (homoskedasticity). Assuming also distributional correctness for the errors (normality), one obtains finite-sample correct tests and confidence intervals.

The analogous 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 $(\vec{\boldsymbol{x}}_i', y_i)$ of the data matrix $(\boldsymbol{X}, \boldsymbol{y})$ are i.i.d. samples from a joint multivariate distribution that has moments to some order. Thus sandwich-based theory provides asymptotically correct inference that is **assumption-lean** or **model-robust**; linear models theory provides finite-sample correct inference that is **assumption-laden** or **model-trusting**. The question arises what sandwich-based inference is about: When no model is assumed, what are the parameters, and what is their meaning?

Answering these and related questions is a first goal of the present article. An established answer is that parameters can be interpreted as statistical functionals $\boldsymbol{\beta}(\boldsymbol{P})$ defined on a large nonparametric class of joint distributions $\boldsymbol{P} = \boldsymbol{P}(d\vec{\boldsymbol{x}}, dy)$ through best approximation (Section 3). The sandwich estimator produces then asymptotically correct standard errors for the slope functionals $\beta_j(\boldsymbol{P})$ (Section 5). The question of the meaning of slopes in the presence of nonlinearity will be answered with proposals involving case-wise and pairwise slopes (Section 8).

A second goal of this article is to discuss the role of the regressors when they are random. Assumption-lean asymptotic theory treats the regressors as random, whereas assumption-laden theory tends to condition on them and treat them as fixed. The justification for conditioning on regressors derives from the ancillarity principle. It will be shown that in an assumption-lean theory the principle is violated: population parameters depend on the distribution of the regressors (Section 4), and the nonlinearity "conspires" with the randomness of the regressors to generate a contribution to the standard errors (Section 5).

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 assumption-lean 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. Thus both may be called **assumption-lean** or **model-robust estimators** (Section 6).

A fourth goal of this article is to practically (Section 2) and theoretically (Section 9) compare assumption-lean and usual estimators. We define a ratio of asymptotic variances — "***RAV***" for short — that describes the discrepancies between the two standard errors in the asymptotic limit. If there exists a discrepancy, $\boldsymbol{RAV} \neq 1$, it is assumption-lean estimators (sandwich or x-y bootstrap) that are asymptotically correct, and the usual standard error is indeed asymptotically

incorrect. The ***RAV*** can range from 0 to ∞ under scenarios that give insight into the nature of model deviations that 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 the presence of model violations that invalidate the usual standard error of a specific coefficient. Although the result can be called a "misspecification test," it is more usefully viewed as a discrepancy test for standard errors, separately for each coefficient (Section 10).

A final goal is to briefly discuss issues with the sandwich estimator: When the model is correct, the sandwich estimator can be inefficient. We will additionally point out that it is also non-robust in the sense of sensitivity to outlying observations. On this topic we will not have more to offer than suggestions.

A feature of the present article is that it makes strong use of regressor adjustment (Section 7) which permits the representation of a multiple regression coefficient as a simple regression coefficient on its adjusted regressor. This fact allows the analysis to be undertaken for one regression coefficient at a time.

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. The linear model is used to allow explicit calculations, but most conclusions extend to a large class of moment estimators, hinted at in paragraphs with the header "Generalizations," which readers may initially skip.

Readers familiar with the sandwich estimator may skim the article for appearances of the nonlinearity η, which is the aspect of this work that is least known. Readers may also browse the tables and figures and read associated sections that seem most germane. Important notations are shown in boxes for reference.

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." Wasserman's (2011) wide-ranging discussion calls for "Low Assumptions, High Dimensions." Davies' (2014) book elaborates the idea of adequate models for a given sample size.

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

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	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.

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

3.1 Targets of Estimation

To make standard errors meaningful it is necessary to first define targets of estimation. As mentioned in the introduction, parameters of generative models are reinterpreted as statistical functionals that are well-defined for a large nonparametric class of data distributions. In an assumption-lean population framework for linear 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 have a joint distribution,

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

whose second moments exist and whose regressors have a full rank covariance matrix. We write

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

for the *column* random vector consisting of the regressor variables, with a constant 1 prepended to accommodate an intercept. Values of the random vector $\vec{\boldsymbol{X}}$ will be denoted by lower case $\vec{\boldsymbol{x}} = (1, x_1, ..., x_p)'$. We write the joint distribution of $(Y, \vec{\boldsymbol{X}})$, the marginal distribution of $\vec{\boldsymbol{X}}$, and the conditional distribution of Y given $\vec{\boldsymbol{X}}$, respectively, as $\boldsymbol{P} = \boldsymbol{P}(\mathrm{d}y, \mathrm{d}\vec{\boldsymbol{x}})$, $\boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}})$, and $\boldsymbol{P}(\mathrm{d}y \,|\, \vec{\boldsymbol{x}})$, or alternatively as $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$, $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, and $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. Nonsingularity of the $p \times p$ regressor covariance matrix is equivalent to nonsingularity of the $(p+1)\times(p+1)$ matrix $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']$.

Due to the prepended intercept coordinate 1, the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is degenerate in $\mathbb{R}^{p+1}$. In addition, there may arise nonlinear degeneracies if multiple regressors are functions of one underlying independent variable, as in polynomial or B-spline regression, or if product interactions are included. These cases of degeneracies are permitted as long as $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']$ remains non-singular.

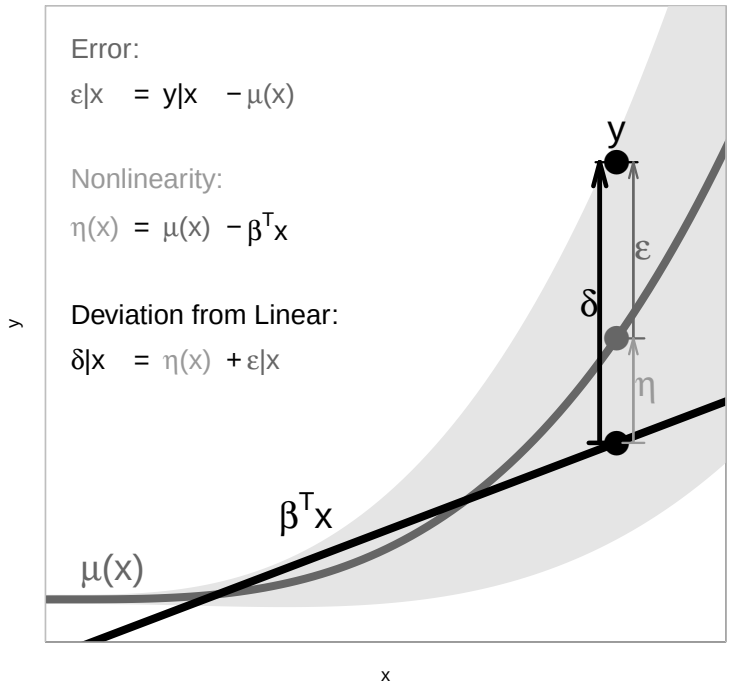

FIG 1. *Illustration of the decomposition* (13).

We write any function $f(X_1, ..., X_p)$ of the regressors as $f(\vec{\boldsymbol{X}})$ as the prepended constant 1 is irrelevant. The following functions of $\vec{\boldsymbol{X}}$ are special:

- ***The best*** $L_2(\boldsymbol{P})$ ***approximation*** to Y, $\mu(\vec{\boldsymbol{X}})$, is the conditional expectation of Y given $\vec{\boldsymbol{X}}$:

$$\mu(\vec{\boldsymbol{X}}) \;:=\; \mathrm{argmin}_{f(\vec{\boldsymbol{X}})\in L_2(\boldsymbol{P})} \boldsymbol{E}[(Y - f(\vec{\boldsymbol{X}}))^2] \;=\; \boldsymbol{E}[\,Y \,|\, \vec{\boldsymbol{X}}\,]\,. \tag{4}$$

  Also called the "response surface," it is ***not*** assumed to be linear in $\vec{\boldsymbol{X}}$.

- ***The best population linear approximation*** to Y is $l(\vec{\boldsymbol{X}}) = \boldsymbol{\beta}'\vec{\boldsymbol{X}}$ whose coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ are given by

$$\boldsymbol{\beta}(\boldsymbol{P}) := \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \boldsymbol{E}[(Y - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2] \qquad = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}Y\,] \tag{5}$$

$$= \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \boldsymbol{E}[(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2] = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})\,] \tag{6}$$

The right-most expressions in (5) and (6) follow from the normal equations:

$$\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\,\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}Y] \;=\; \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\,\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})] \;=\; \boldsymbol{0}. \tag{7}$$

We use the shorthand "population coefficients" for $\boldsymbol{\beta}(\boldsymbol{P})$ and "population approximation" for $\boldsymbol{\beta}(\boldsymbol{P})'\vec{\boldsymbol{X}}$, omitting "linear" and "OLS". We will often write $\boldsymbol{\beta}$, omitting the argument $\boldsymbol{P}$, when it is clear that $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$. The population coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ form a vector ***statistical functional*** defined for a large class of joint data distributions $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$.

Generalizations:

- An assumption-lean interpretation of the maximum likelihood (ML) method is as follows: Given a regression model $p(y \mid \vec{\boldsymbol{x}}; \boldsymbol{\theta})$ define a statistical functional by minimization,

$$\boldsymbol{\theta}(\boldsymbol{P}) = \text{argmin}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{P}}[-\log p(Y|\vec{\boldsymbol{X}}; \boldsymbol{\theta})], \tag{8}$$

  or by solving the associated moment conditions/estimating equations,

$$\boldsymbol{E}_{\boldsymbol{P}}[\partial/\partial\boldsymbol{\theta} \log p(Y|\vec{\boldsymbol{X}}; \boldsymbol{\theta})] = \mathbf{0}. \tag{9}$$

  Under mild regularity conditions we have $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{\theta}_0$ if the actual conditional data distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ has density $p(y \mid \vec{\boldsymbol{x}}; \boldsymbol{\theta}_0)$. The point is, however, that $\boldsymbol{\theta}(\boldsymbol{P})$ is defined for a large class of data distributions outside of the model $p(y \mid \vec{\boldsymbol{x}}; \boldsymbol{\theta})$.
- In the assumption-lean view of the ML method the two-fold role of the model is 1) to provide a heuristic for a loss function

$$\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) \;=\; -\log p(y \mid \vec{\boldsymbol{x}}; \boldsymbol{\theta}),$$

  and 2) to act as an approximation to the actual conditional data distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. (An early adopter of this point of view is Kent (1982).) Generalizing this view one may define statistical functionals from loss functions $\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$ that are not necessarily derived from models:

$$\boldsymbol{\theta}(\boldsymbol{P}) \;=\; \text{argmin}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{P}}[\mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})]. \tag{10}$$

  (Note that such $\mathcal{L}$ are not loss functions in the sense of Wald's decision theory, but the terminology has become entrenched in machine learning.)
- The minimization problem (10) is usually solved in terms of stationarity conditions that amount to moment conditions for $\boldsymbol{\psi}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) = \partial_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})$:

$$\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] \;=\; \mathbf{0}. \tag{11}$$

  It is natural to generalize further and define statistical functionals as solutions to moment conditions (11) where $\boldsymbol{\psi}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$ is not required to be the gradient of any loss function; in particular it need not be the score function of a likelihood. A seminal work that inaugurated asymptotic theory for very general moment conditions is by Huber (1967). For OLS, (11) specializes to the normal equations (7) as the moment function for the slopes is

$$\boldsymbol{\psi}_{OLS}(\boldsymbol{\beta}; y, \vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}}\vec{\boldsymbol{x}}'\boldsymbol{\beta} - \vec{\boldsymbol{x}}y, \tag{12}$$

- An extension to situations where the number of moment conditions (the dimension of $\boldsymbol{\psi}$) is larger than the dimension of $\boldsymbol{\theta}$ is provided by the Generalized Method of Moments (GMM, Hansen 1982). It is intended for causal inference based on numerous instrumental variables.
- A generalization of moment conditions to clustered data with intra-cluster dependence is provided by Generalized Estimating Equations (GEE, Liang and Zeger 1986). This approach, however, is not framed in terms of statistical functionals of joint $(Y, \vec{\boldsymbol{X}})$ distributions; it is rather a "fixed-$\boldsymbol{X}$" approach that assumes well-specification of the mean function while allowing misspecification of variance and intra-cluster dependence.

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.

3.2 The Canonical Noise-Nonlinearity Decomposition and its Properties

We continue with the OLS case for the sake of simplicity, explicit formulas and direct insights. The response Y has the following canonical decompositions:

$$
\begin{aligned}
Y &= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \underbrace{(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\vec{\boldsymbol{X}})}_{\eta(\vec{\boldsymbol{X}})} + \underbrace{(Y - \mu(\vec{\boldsymbol{X}}))}_{\epsilon} \\
&= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \underbrace{\eta(\vec{\boldsymbol{X}}) + \epsilon}_{\delta} \\
&= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \delta
\end{aligned}
\tag{13}
$$

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. The following list contains mutual relations between the regressors and the components of the canonical decompositions, as well as some further definitions:

- ***Medium-sense orthogonality of noise***: The noise ϵ satisfies $\epsilon \perp L_2(\boldsymbol{P}_{\vec{\boldsymbol{X}}})$:

$$
\boldsymbol{E}[\,\epsilon\, f(\vec{\boldsymbol{X}})\,] \;=\; 0 \quad \forall f(\vec{\boldsymbol{X}}) \in L_2(\boldsymbol{P}_{\vec{\boldsymbol{X}}}),
\tag{14}
$$

  which is equivalent to conditional centering, $\boldsymbol{E}[\,\epsilon\,|\vec{\boldsymbol{X}}] \overset{\boldsymbol{P}}{=} 0$. It is ***not independent of*** $\vec{\boldsymbol{X}}$, which we would call "strong sense orthogonal" because of the equivalence to $L_2(\epsilon) \perp L_2(\boldsymbol{P}_{\vec{\boldsymbol{X}}})$.

- ***Weak-sense orthogonalities***: η, ϵ, $\delta \perp \vec{\boldsymbol{X}}$, that is,

$$
\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\eta\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\epsilon\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\delta\,] \;=\; \boldsymbol{0}.
\tag{15}
$$

  The first, $\eta \perp \vec{\boldsymbol{X}}$, holds because by (6) η is the population residual of the OLS linear regression of $\mu(\vec{\boldsymbol{X}})$ on $\vec{\boldsymbol{X}}$; the second, $\epsilon \perp \vec{\boldsymbol{X}}$, follows from (14); finally, $\delta \perp \vec{\boldsymbol{X}}$ because $\delta = \eta + \epsilon$.

- ***Marginal centering***, unconditional, is a special case of (15) due to the inclusion of an intercept in $\vec{\boldsymbol{X}}$:

$$
\boldsymbol{E}[\,\eta\,] = \boldsymbol{E}[\,\epsilon\,] = \boldsymbol{E}[\,\delta\,] = 0.
\tag{16}
$$

- ***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 existing conditional variances. Define:

$$
\sigma^2(\vec{\boldsymbol{X}}) := \boldsymbol{V}[\,\epsilon\,|\vec{\boldsymbol{X}}] \;=\; \boldsymbol{E}[\,\epsilon^2\,|\,\vec{\boldsymbol{X}}] \;\overset{\boldsymbol{P}}{<}\; \infty.
\tag{17}
$$

- ***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:

$$m^2(\vec{\boldsymbol{X}}) \;:=\; \boldsymbol{E}[\,\delta^2 \,|\, \vec{\boldsymbol{X}}\,] \;=\; \eta^2(\vec{\boldsymbol{X}}) + \sigma^2(\vec{\boldsymbol{X}}). \tag{18}$$

  The decomposition follows from $\delta = \eta + \epsilon$ and $\epsilon \perp \eta(\vec{\boldsymbol{X}})$ due to (14).

- ***Mean squared functionals***:

$$\begin{array}{lcll} \eta^2(\boldsymbol{P}) & := & \boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})\,], & \text{mean squared nonlinearity,} \\ \sigma^2(\boldsymbol{P}) & := & \boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\,] = \boldsymbol{E}[\,\epsilon^2\,], & \text{mean noise variance,} \\ m^2(\boldsymbol{P}) & := & \boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,], & \text{mean or plain MSE.} \end{array} \tag{19}$$

  All expectations, except for $\boldsymbol{E}[\,\epsilon^2\,]$, are w.r.t. $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. From (18) follows

$$m^2(\boldsymbol{P}) = \eta^2(\boldsymbol{P}) + \sigma^2(\boldsymbol{P}). \tag{20}$$

- ***Well-specification*** can be expressed to first order as $\eta^2(\boldsymbol{P}) \overset{\boldsymbol{P}}{=} 0$ and to second order as $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} \sigma^2(\boldsymbol{P}) = \text{const}$. These do not imply well-specification w.r.t. Gaussianity of the error distribution.

In what follows one must keep in mind that the nonlinearity $\eta(\vec{\boldsymbol{X}})$ is weakly orthogonal to the regressors, that is, centered and uncorrelated with all X_j.

3.3 Error Terms and Random Regressors: Uncorrelated versus Independent

The term "error" has been carefully avoided so far. The following brief digression relates the notion of "error term" to the present framework. If a response Y is modeled as $Y = f(\vec{\boldsymbol{X}}; \boldsymbol{\theta}) + e$, where $\vec{\boldsymbol{X}}$ is random, one has to specify a stochastic relation between $\vec{\boldsymbol{X}}$ and e. If it is reasonable to assume that the errors are unassociated with the regressors, three possibilities exist:

Definitions and Lemma 3.3:

- ***Weak-sense error terms****: e and $\vec{\boldsymbol{X}}$ are orthogonal, $\boldsymbol{E}[e\,\vec{\boldsymbol{X}}] = \boldsymbol{0}$.*
  Such errors permit both nonlinearities and heteroskedasticities, hence misspecification to first and second order is not meaningful.
- ***Medium-sense error terms****: e and $L_2(\vec{\boldsymbol{X}})$ are orthogonal.*
  Such errors permit heteroskedasticities but not nonlinearities, hence misspecification to first order is meaningful, but not to second order.
- ***Strong-sense error terms****: e and $\vec{\boldsymbol{X}}$ are independent.*
  Such errors exclude both nonlinearities and heteroskedasticities, hence misspecification to first and second order are both meaningful.

White (1980b) navigates the distinction between weak- and strong-sense error terms as follows: In his Section 2 (p. 818) he assumes weak-sense error terms without noting that these allow inclusion not only of heteroskedasticities but nonlinearities as well. In his Section 3 (p. 824) in the context of a heteroskedasticity test, he notices that this is the same test he proposed in White (1980a) for nonlinearity. His null hypothesis implies strong-sense error terms which preclude both nonlinearity and heteroskedasticity.

The discussion in this subsection has been about the stochastic relation between random regressors and error terms in the population. It is unrelated to the assumption of i.i.d. errors among observations when the regressors are fixed.

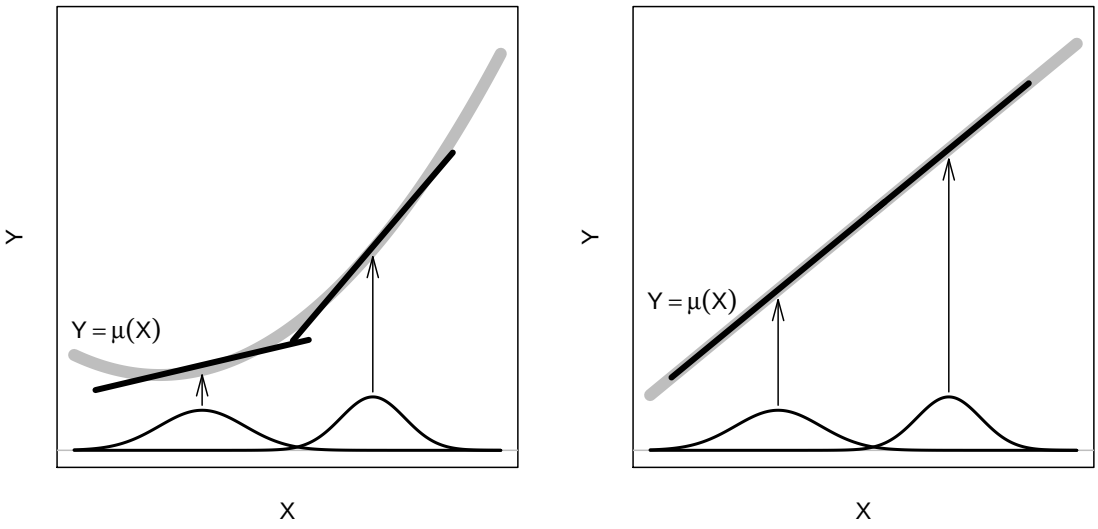

FIG 2. *Illustration of the dependence of the population OLS solution on the marginal distribution of the regressors: The left figure shows dependence in the presence of nonlinearity; the right figure shows independence in the presence of linearity.*

4. NON-ANCILLARITY OF THE REGRESSOR DISTRIBUTION

4.1 The Breakdown of the Ancillarity Argument

Conditioning on the regressors when they are random has historically been justified with the ancillarity principle. The argument applies to any regression model rendered in the following form:

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

referring to the model densities of $\boldsymbol{P}_{\vec{\boldsymbol{X}},Y}$, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, respectively. The parameter of interest is $\boldsymbol{\theta}$ while the regressor density $p(\vec{\boldsymbol{x}})$ acts 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 $p(\vec{\boldsymbol{x}})$. (For a fuller definition of ancillarity see Appendix B.) This logic is valid if the conditional model $p(y \,|\, \vec{\boldsymbol{x}}; \boldsymbol{\theta})$ is correct. The following proposition describes for linear models the ways in which ancillarity is broken if the model is an approximation rather than a truth.

Proposition 4.1:

- *Among distributions $\boldsymbol{P}$ that share the conditional expectation $\mu(\vec{\boldsymbol{x}})$, the functional $\boldsymbol{\beta}(\boldsymbol{P})$ depends on the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ if and only if $\mu(\vec{\boldsymbol{x}})$ is nonlinear.*
- *Among distributions $\boldsymbol{P}$ that share the conditional variance $\sigma^2(\vec{\boldsymbol{x}})$, the functional $\sigma^2(\boldsymbol{P})$ depends on the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ if and only if $\sigma^2(\vec{\boldsymbol{x}})$ is non-constant (heteroskedastic).*

(These are loose statements; see Appendix D.1 for more precision.) The first part of the proposition is best explained with a graphical illustration: Figure 2 shows single regressor situations with a nonlinear and a linear mean function,

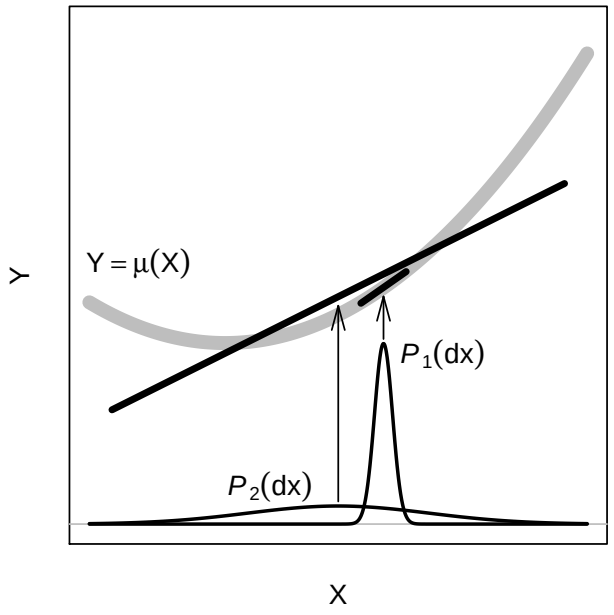

FIG 3. *Illustration of the interplay between regressors' high-density range and nonlinearity: Over the small range of $\boldsymbol{P}_1$ the nonlinearity will be undetectable and immaterial for realistic sample sizes, whereas over the extended range of $\boldsymbol{P}_2$ the nonlinearity is more likely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

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. (This observation appears first in White (1980a, p. 155f); to see the correspondence, identify Y with his $g(Z) + \epsilon$.)

Ancillarity of regressors is sometimes informally explained as the regressor distribution being independent of, or unaffected by, the parameters of interest. 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 variable, 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 regressor distributions can become increasingly complex for larger regressor dimensions or, worse, as $p \to \infty$. Differences in estimated parameter values often become visible in meta-analyses and may be interpreted as "parameter heterogeneity." The source of this heterogeneity may be differences in covariate distributions combined with nonlinearities relative to the fitted model.

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

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$,		parameter vector	$((p+1)\times 1)$
$\boldsymbol{Y} = (Y_1, ..., Y_N)'$,		response vector	$(N\times 1)$
$\boldsymbol{X}_j = (X_{1,j}, ..., X_{N,j})'$,		j'th regressor vector	$(N\times 1)$
$\boldsymbol{X} = [\mathbf{1}, \boldsymbol{X}_1, ..., \boldsymbol{X}_p]$	$= \begin{bmatrix} \vec{\boldsymbol{X}}_1{}' \\ ..... \\ ..... \\ \vec{\boldsymbol{X}}_N{}' \end{bmatrix}$,	regressor matrix with intercept	$(N\times(p+1))$
$\boldsymbol{\mu} = (\mu_1, ..., \mu_N)'$,	$\mu_i = \mu(\vec{\boldsymbol{X}}_i) = \boldsymbol{E}[Y\|\vec{\boldsymbol{X}}_i]$,	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	$(N\times 1)$
$\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_N)'$,	$\sigma_i = \sigma(\vec{\boldsymbol{X}}_i) = \boldsymbol{V}[Y\|\vec{\boldsymbol{X}}_i]^{1/2}$,	conditional sdevs	$(N\times 1)$
$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)'$	$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$,	parameter estimates	$((p+1)\times 1)$
$\boldsymbol{r} = (r_1, ..., r_N)'$	$= \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$,	sample residuals	$(N\times 1)$

TABLE 3
Random variable notation for i.i.d. observational data.

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 excellent, hence the model 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. This is a general issue that holds even in the most successful theories, those of physics, which at this point in history have limited ranges of validity as well.

5. OBSERVATIONAL DATASETS, ESTIMATION, AND CLTS

We turn from populations to estimation from i.i.d. data. We sacrifice the generality that is common in econometrics and trade it for simplicity. White (1980b), for example, assumes observations to be "independent not (necessarily) identically distributed", and Hansen (1982) assumes them stationary and ergodic. The goal is to describe how the sampling variability of estimates decomposes according to two sources, noise and nonlinearity, with emphasis on the latter.

5.1 Observational Datasets and Estimation

Assume data consisting of i.i.d. cases/observations $(Y_i, X_{i,1}, ..., X_{i,p})$ drawn from a joint multivariate distribution $\boldsymbol{P}(\mathrm{d}y, \mathrm{d}x_1, ..., \mathrm{d}x_p)$ $(i = 1, 2, ..., N)$, and stack them as in Table 3. The definitions of η, ϵ and δ translate to N-vectors:

$$\boldsymbol{\eta} \;=\; \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} \;=\; \boldsymbol{Y} - \boldsymbol{\mu}, \qquad \boldsymbol{\delta} \;=\; \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \;=\; \boldsymbol{\eta} + \boldsymbol{\epsilon}. \tag{21}$$

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 $\boldsymbol{X}_j$ in the sample.

Writing $\langle \cdot, \cdot \rangle$ for the usual Euclidean inner product on $\mathbb{R}^N$, we have in general

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

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 (15).

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

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\tilde{\boldsymbol{\beta}}} \|\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\,. \tag{22}$$

Because we are not conditioning on $\boldsymbol{X}$, randomness of $\hat{\boldsymbol{\beta}}$ stems from $\boldsymbol{Y}$ as well as $\boldsymbol{X}$. The sample residual vector $\boldsymbol{r} = \boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$, which arises from $\hat{\boldsymbol{\beta}}$, is distinct from the population residual vector $\boldsymbol{\delta} = \boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}$, 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.

5.2 Decomposition of OLS Estimates According to Noise and Nonlinearity

In $\boldsymbol{X}$-conditional linear models theory, the target of estimation $\boldsymbol{\beta}(\boldsymbol{X})$ is what we may call the "conditional parameter":

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

Unconditionally, $\boldsymbol{\beta}(\boldsymbol{X})$ is a random variable, hence is generally not the target of estimation, which is $\boldsymbol{\beta}(\boldsymbol{P})$ in a random-$\boldsymbol{X}$ theory. In what follows we analyze the relationship between $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$, $\boldsymbol{\beta}(\boldsymbol{X})$ and $\boldsymbol{\beta}(\boldsymbol{P})$. It will be shown that the unconditional true standard error permits a Pythagorean decomposition into contributions due to noise and due to nonlinearity, both of order $1/\sqrt{N}$, according to

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\boldsymbol{X})) + (\boldsymbol{\beta}(\boldsymbol{X}) - \boldsymbol{\beta}). \tag{23}$$

Definition and Lemma 5.2: *Define "Estimation Offsets" (EOs) as follows:*

$$\boxed{\begin{array}{lrcl} \textit{Total EO} & := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} & = & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\delta}, \\ \textit{Noise EO} & := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\boldsymbol{X}) & = & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}, \\ \textit{Model Deviation EO} & := \boldsymbol{\beta}(\boldsymbol{X}) - \boldsymbol{\beta} & = & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}. \end{array}} \tag{24}$$

The right hand equalities follow from the decompositions (21), $\boldsymbol{\epsilon} = \boldsymbol{Y}-\boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\delta}=\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}$, and these facts:

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

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

Generalizations: The three *EO*s can be defined for general moment estimators in regressor-response data. Here are the moment conditions that define $\boldsymbol{\theta}$, $\boldsymbol{\theta}(\boldsymbol{X})$ and $\hat{\boldsymbol{\theta}}$, respectively:

$$\boxed{\begin{array}{ll} \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{P}): & \boldsymbol{E}[\,\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})\,] = \boldsymbol{0}, \\ \boldsymbol{\theta}(\boldsymbol{X}): & \frac{1}{N}\sum_i \boldsymbol{E}[\,\boldsymbol{\psi}(\boldsymbol{\theta}; Y_i, \vec{\boldsymbol{X}}_i)\,|\,\vec{\boldsymbol{X}}_i] = \boldsymbol{0}, \\ \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{P}}): & \frac{1}{N}\sum_i \boldsymbol{\psi}(\boldsymbol{\theta}; Y_i, \vec{\boldsymbol{X}}_i) = \boldsymbol{0}. \end{array}} \tag{25}$$

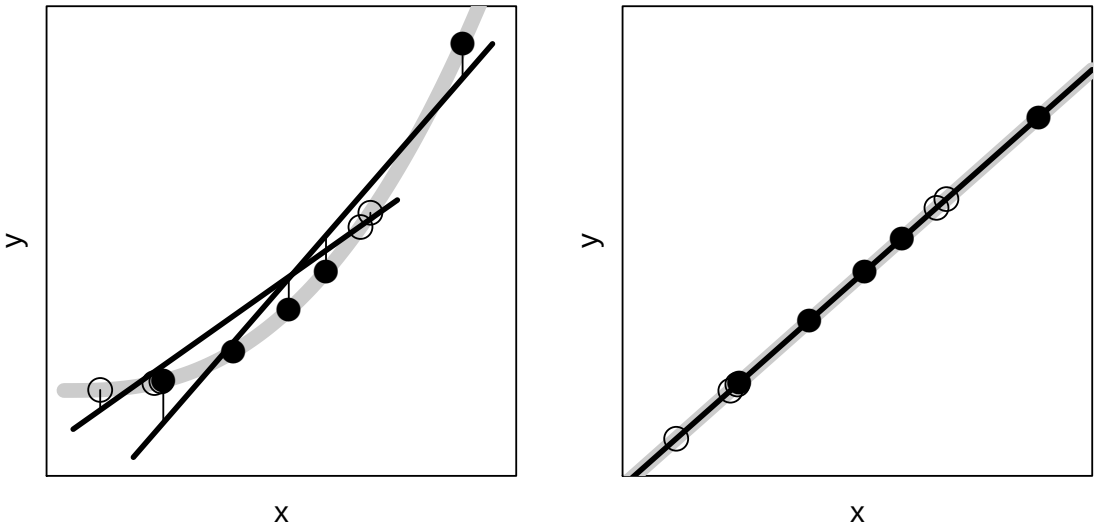

FIG 4. *Noise-less Response: The filled and the open circles represent two "datasets" from the same population. The x-values are random; the y-values are a deterministic function of x: $y = \mu(x)$ (shown in gray).*
Left: The true response $\mu(x)$ is nonlinear; the open and the filled circles have different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

These specialize to OLS for $\boldsymbol{\psi}_{OLS}(\boldsymbol{\beta}; y, \vec{\boldsymbol{x}})$ (12). The clean attributions of the three EOs to $\boldsymbol{\delta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ according to the above lemma exists only when $\boldsymbol{\psi}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}})$ is affine in y for fixed $\vec{\boldsymbol{x}}$ and $\boldsymbol{\theta}$. Just the same we can define:

$$
\boxed{
\begin{array}{lcl}
\textit{Total EO} & := & \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \\
\textit{Noise EO} & := & \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{X}), \\
\textit{Model Deviation EO} & := & \boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}.
\end{array}
} \tag{26}
$$

The role of $\boldsymbol{\theta}(\boldsymbol{X})$ is to reflect the systematic part of the moment condition. It is a function of a sample of regressor tuples $\vec{\boldsymbol{X}}_i$ and the conditional distributions $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}_i}$ of Y at these tuples. Intuitively, it is about an idealized situation where at each $\vec{\boldsymbol{X}}_i$ one observes not one but infinitely many response values Y_i.

5.3 Random X and Model Deviation as a Source of Sampling Variation

From the point of view of fixed-$\boldsymbol{X}$ linear models theory the model deviation EO, $\boldsymbol{\beta}(\boldsymbol{X}) - \boldsymbol{\beta}$, is a bias. In truth it is random vector when $\boldsymbol{X}$ is random, and as such it is a source of sampling variability. This fact is best illustrated with a noise-free situation: Consider a response that is a deterministic but nonlinear function of the regressors, $Y = \mu(\vec{\boldsymbol{X}})$, so that in a sample $\boldsymbol{\epsilon} = \boldsymbol{0}$ but $\boldsymbol{\eta} \neq \boldsymbol{0}$, and hence $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{X}) - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}$, where $\boldsymbol{X}$ as well as $\boldsymbol{\eta}$ are both random. If $\mu(\vec{\boldsymbol{x}})$ were linear, this term would vanish due to $\boldsymbol{\eta} = \boldsymbol{0}$ and $\hat{\boldsymbol{\beta}}$ would be constant. Thus for sampling variability from this source to exist, $\boldsymbol{X}$ needs to be random and $\mu(\vec{\boldsymbol{X}})$ needs to be nonlinear — the "conspiracy" in the title of the article.

Figure 4 illustrates the situation with a single-regressor example by showing the OLS lines fitted to two "datasets" consisting of $N = 5$ regressor values each. The

random differences between datasets cause the fitted line to exhibit sampling variability under nonlinearity (left hand plot), which is absent under linearity (right hand plot). Comparing with Figure 2, we realize that the effect illustrated in both is the same, but it is shown for different populations in Figure 2 and for different datasets in Figure 4. Thus nonlinearity has two fundamental effects: (1) the population parameter $\boldsymbol{\beta}(\boldsymbol{P})$ becomes dependent on the regressor distribution, and (2) the "conditional parameter" $\boldsymbol{\beta}(\boldsymbol{X})$ exhibits sampling variability.

Generalizations: For general moment estimators, the "conditional parameter" $\boldsymbol{\theta}(\boldsymbol{X})$ of (25) is generally again a random vector. The above discussion motivates generalizing the notion of "well-specification" to moment conditions for a given joint distribution or, more precisely, for the conditional distributions $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$:

Definition: *The moment conditions* $\boldsymbol{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})] = \mathbf{0}$ *are well-specified for* $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ *if* $\boldsymbol{\theta}(\boldsymbol{X}) \overset{\boldsymbol{P}}{=} \boldsymbol{\theta}(\boldsymbol{P})$ *is constant across regressor matrices* $\boldsymbol{X}$ *that provide identifiability of* $\boldsymbol{\theta}$ *through the conditional moment conditions that define* $\boldsymbol{\theta}(\boldsymbol{X})$*:* $\sum_i \boldsymbol{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}}_i \,|\, \vec{\boldsymbol{X}}_i\,]) = \mathbf{0}$.

This definition characterizes joint distributions $\boldsymbol{P}$ for which the same unconditional parameter $\boldsymbol{\theta}(\boldsymbol{P})$ is estimated conditionally on $\boldsymbol{X}$ irrespective of $\boldsymbol{X}$. Equivalently, for such $\boldsymbol{P}$ the model deviation EO vanishes: $\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}(\boldsymbol{P}) \overset{\boldsymbol{P}}{=} \mathbf{0}$. This generalizes the situation illustrated by the right hand plot of Figure 4. Well-specification in this sense is a parameter-dependent notion. In OLS, well-specification for the slopes speaks to the linearity of the response surface as a function of the regressors, whereas well-specification for the variance speaks to the constancy of the conditional noise variance.

Fixed-$\boldsymbol{X}$ theories of regression, such as linear models theory, necessarily assume well-specification in the sense of the above definition. Their problem is that their only source of sampling variability is the noise EO $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{X})$. The "remedy" of fixed-$\boldsymbol{X}$ theories is to call for model diagnostics and declare a model and its inferences to be invalid if model deviations are detected. If there exist model deviations that cause $\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta} \neq \mathbf{0}$ but remain undetected in a particular data analysis, they will be erroneously treated as arising from noise, and statistical inference may consequently be invalidated (Section 9.4). This mistreatment occurs in the residual bootstrap which assumes the residuals to originate from exchangeable noise. Asymptotically correct treatment is provided by the sandwich estimator and the x-y bootstrap, even in noise-free nonlinear situations. The justifications derive from central limit theorems which are described next.

5.4 Assumption-Lean Central Limit Theorems

Proposition 5.4: *The three EOs follow CLTs for fixed* p *as* $N \to \infty$*:*

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

The proofs are standard. — Note that the contribution of the nonlinearity is of the same order $1/\sqrt{N}$ as the contribution of the noise. The CLTs are shown in terms of the decomposition (18), $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$, but by (17,18) $m^2(\vec{\boldsymbol{X}})$ can be replaced by δ^2 and $\sigma^2(\vec{\boldsymbol{X}})$ by ϵ^2, which will be used later:

$$(28) \quad \boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] \;=\; \boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'], \qquad \boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] \;=\; \boldsymbol{E}[\,\epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'].$$

Consider some special cases:

- **First order well-specification:** $\eta(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$.

  $$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\,\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\;\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\right)$$

  The sandwich form is solely due to heteroskedasticity.

- **Deterministic nonlinear response:** $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$.

  $$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\,\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\;\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\right)$$

  The sandwich form is solely due to the nonlinearity and randomness of $\boldsymbol{X}$.

- **First and second order well-specification:** $\eta(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0, \sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} \sigma^2(\boldsymbol{P})$.

  $$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \sigma^2\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\right).$$

  This non-sandwich form is asymptotically valid without Gaussian errors.

Generalizations: The CLT for $\hat{\boldsymbol{\beta}}$ is a special case of assumption-lean CLTs for moment conditions due to Huber (1967). For a generic vector moment condition $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(Y, \vec{\boldsymbol{X}}; \boldsymbol{\theta})] = \mathbf{0}$ that defines a statistical functional $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{P})$ (Section 3.1) with plug-in estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{P}})$, there hold under technical conditions the following CLTs, where $\boldsymbol{\Lambda}(\boldsymbol{\theta}) := \partial_{\boldsymbol{\theta}}\boldsymbol{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})]$ is a Jacobian, $\dim(\boldsymbol{\psi}) \times \dim(\boldsymbol{\theta})$:

$$(29) \quad \boxed{\begin{aligned}
\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \;&\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{V}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})]\;\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}\right) \\
\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\boldsymbol{X})) \;&\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{E}[\boldsymbol{V}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,]\;\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}\right) \\
\sqrt{N}\,(\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}) \;&\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\,\boldsymbol{V}[\boldsymbol{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\boldsymbol{X}})\,|\,\vec{\boldsymbol{X}}]\,]\;\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}\right)
\end{aligned}}$$

The first is Huber's (1967) result. All three specialize to OLS (27) for $\boldsymbol{\psi}_{OLS}$ (12), and to assumption-lean ML for $\boldsymbol{\psi}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) = -\partial_{\boldsymbol{\theta}} \log p(y|\vec{\boldsymbol{x}}; \boldsymbol{\theta})$. It is natural that the asymptotic variances of the EOs are related according to the identity

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

where on the right side the first summand relates to the noise EO and the second to the model deviation EO.

6. THE SANDWICH ESTIMATOR 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. This is intuitively unsurprising as they both estimate the same asymptotic variance, that of the first CLT in Proposition 5.4. A closer connection between them will be established below.

6.1 The Plug-In Sandwich Estimator of Asymptotic Variance

According to Proposition 5.4 and (28) the asymptotic variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\boldsymbol{AV}[\hat{\boldsymbol{\beta}}] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}'\,]^{-1}\,\boldsymbol{E}[\,\delta^2\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}'\,]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}'\,]^{-1}. \tag{30}$$

The sandwich estimator is then the plug-in version of (30) where δ^2 is replaced by residuals and population expectations $\boldsymbol{E}[...]$ by sample means $\hat{\boldsymbol{E}}[...]$:

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

where $\boldsymbol{D}(\boldsymbol{r})^2$ is the diagonal matrix with squared residuals $r_i^2 = (Y_i - \vec{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})^2$ in the diagonal. With this notation the simplest and original form of the sandwich estimator of asymptotic variance can be written as follows (White 1980a):

$$\begin{aligned} \hat{\boldsymbol{AV}}_{sand} \;&:=\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}'\,]^{-1}\,\hat{\boldsymbol{E}}[\,r^2\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}'\,]\,\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}'\,]^{-1} \\ &=\; N\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\,(\boldsymbol{X}'\boldsymbol{D}(\boldsymbol{r})^2\,\boldsymbol{X})\,(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned} \tag{31}$$

This estimator is asymptotically consistent. The sandwich standard error estimate for the j'th regression coefficient is obtained as

$$\hat{\boldsymbol{SE}}_{sand}[\hat{\beta}_j] \;:=\; \tfrac{1}{N^{1/2}}(\hat{\boldsymbol{AV}}_{sand})_{jj}^{\,1/2}. \tag{32}$$

For this simplest version ("HC" in MacKinnon and White (1985)) obvious modifications exist. For one thing, it does not account for the fact that residuals have on average smaller variance than noise. An overall correction factor $(N/(N-p-1))^{1/2}$ in (32) would seem to be sensible in analogy to the linear models estimator ("HC1" ibid.). More detailed modifications have been proposed whereby individual residuals are corrected for their reduced conditional variance according to $\boldsymbol{V}[r_i|\boldsymbol{X}] = \sigma^2(1-H_{ii})$ under homoskedasticity and ignoring nonlinearity ("HC2" ibid.). Further modifications include a version based on the jackknife ("HC3" ibid.) using leave-one-out residuals. An obvious alternative is estimating asymptotic variance with the x-y bootstrap, to which we now turn.

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

To link sandwich 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 resampling *with* replacement from
- M-out-of-N subsampling *without* replacement.

In resampling M can be any $M<\infty$, whereas 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 fro $M \ll N$; 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 resampling is i.i.d. sampling from some distribution, there holds a CLT as the resample size grows, $M \to \infty$. It is immaterial that, in this case, the sampled distribution is the empirical distribution $\boldsymbol{P}_N$ of a given dataset $\{(Y_i, \vec{\boldsymbol{X}}_i)\}_{i=1\ldots N}$, which is frozen of size N as $M \to \infty$.

Proposition 6.2: *For any fixed dataset of size N without exact collinearities, there holds a CLT for the M-of-N bootstrap as $M \to \infty$. Denoting by $\boldsymbol{\beta}^*$ the OLS estimate obtained from a bootstrap resample of size M, we have for $M \to \infty$:*

$$(33)\quad M^{1/2}\,(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\,\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\,\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\right).$$

This is a straight application of the CLT of the previous section to the empirical distribution of the data, where the "meat" of the asymptotic formula is based on the empirical counterpart $r_i^2 = (Y_i - \vec{\boldsymbol{X}}_i'\hat{\boldsymbol{\beta}})^2$ of $\delta^2 = (Y - \vec{\boldsymbol{X}}'\boldsymbol{\beta})^2$. Comparing (31) and (33) leads to the following link between sandwich and bootstrap estimators:

Corollary 6.2: *The sandwich estimator* (31) *is the asymptotic variance estimated by the M-of-N bootstrap in the limit $M \to \infty$ for a fixed sample of size N.*

The sandwich estimator has the advantage that it results 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).

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).

Generalizations: Sandwich estimators of standard error exist for a large class of moment estimators. They are obtained by plug-in into the asymptotic variance given by their CLTs (29):

$$(34)\qquad \hat{\boldsymbol{A}}\hat{\boldsymbol{V}} := \hat{\boldsymbol{\Lambda}}^{-1}\hat{\boldsymbol{V}}[\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}; Y, \vec{\boldsymbol{X}})]\,\hat{\boldsymbol{\Lambda}}'^{-1}, \quad \text{where } \hat{\boldsymbol{\Lambda}} := \hat{\boldsymbol{E}}[\,\partial_{\boldsymbol{\theta}}\,\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}; Y, \vec{\boldsymbol{X}})].$$

Based again on a CLT under the empirical distribution, these sandwich estimators are also the limits of the M-of-N bootstrap when $M \to \infty$ and N is fixed.

7. ADJUSTED REGRESSORS

The following adjustment formulas are standard but will be stated explicitly due to their importance in what follows. They express the slopes of multiple regressions as slopes of simple regressions using adjusted single regressors. The formulas will be used for the interpretation of regression slopes in the presence of nonlinearity (Section 8), the analysis of discrepancies between asymptotically

proper and improper standard errors (Section 9), and a test of discrepancy between the two (Section 10). [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 $Y_{\bullet\text{–}j}$ 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 or $Y_{\bullet\text{–}j}$ on $X_{j\bullet}$:

$$\beta_j \;=\; \frac{\boldsymbol{E}[\,Y_{\bullet\text{–}j}X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}^2]} \;=\; \frac{\boldsymbol{E}[\,Y\,X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}^2]} \;=\; \frac{\boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}})X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}^2]}\,. \tag{35}$$

  The rightmost representation holds because $X_{j\bullet}$ is a function of $\vec{\boldsymbol{X}}$ only which permits conditioning Y on $\vec{\boldsymbol{X}}$ in the numerator.

- **Adjustment in Samples:** Define the sample-adjusted regressor column $\boldsymbol{X}_{j\hat{\bullet}}$ to be the residual vector of the sample regression of $\boldsymbol{X}_j$, used as the response vector, on all other regressors. The response vector $\boldsymbol{Y}$ can be sample-adjusted similarly, and we may denote it by $\boldsymbol{Y}_{\bullet\text{–}j}$ to indicate that $\boldsymbol{X}_j$ is not among the adjustors, which is implicit for $X_{j\bullet}$. (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 $\boldsymbol{Y}$ on $\boldsymbol{X}$ is obtained as the simple regression through the origin of $\boldsymbol{Y}$ or $\boldsymbol{Y}_{\hat{\bullet}\text{–}j}$ on $\boldsymbol{X}_{j\bullet}$:

$$\hat{\beta}_j \;=\; \frac{\langle \boldsymbol{Y}_{\hat{\bullet}\text{–}j}, \boldsymbol{X}_{j\hat{\bullet}}\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2} \;=\; \frac{\langle \boldsymbol{Y}, \boldsymbol{X}_{j\hat{\bullet}}\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,. \tag{36}$$

Generalizations: The adjustment formalism is peculiar to OLS, but formally a weighted version holds for all regressions whose estimates can be computed with iteratively reweighted LS (IRLS) algorithms, including GLMs and robust regressions. Weighted adjustment formulas from IRLS are non-constructive as the algorithm has to be run to find the weights. A form of adjustment could be defined for one-step moment estimators based on a single Newton iteration. One-step estimators starting from $\boldsymbol{\theta} = \boldsymbol{0}$ have the form $\hat{\boldsymbol{\theta}} = -\hat{\boldsymbol{\Lambda}}(\boldsymbol{0})^{-1}\hat{\boldsymbol{E}}[\boldsymbol{\psi}(\boldsymbol{0}; Y, \vec{\boldsymbol{X}})$, which for OLS specializes to the familiar $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{E}}[\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}']^{-1}\hat{\boldsymbol{E}}[Y\,\vec{\boldsymbol{X}}]$. In what follows we will not strive for generality and use OLS instead for qualitative insights.

8. 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: a linear fit in the presence of nonlinearity gives a sense of the direction, up or down, of association between a regressor and the response adjusted for other regressors. (If the sole purpose is response prediction, well-specification is not the goal either; it is rather trading off nonlinearity against noise over the regressor range.)

The issue is that, in the presence of nonlinearity, slopes lose their usual interpretation: β_j is no longer the average difference in Y associated with a unit difference in X_j at fixed levels of all other X_k. 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. Regression slopes can then be interpreted as weighted averages of "case-wise" and "pairwise" slopes in a sense to be made precise. 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." ("Linearly" will be dropped in what follows.)

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

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

By (35) and (36), the population slopes and their estimates are, respectively,

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

Slope interpretation will be based on the following devices:

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

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

    so b and w where are case-wise slopes and case-wise weights, respectively.
  - **pairwise slopes**:

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

    so b and w are pairwise slopes and weights, respectively, and (x, y) and (x', y') are two independent identically distributed copies of the adjusted regressor-response distribution.

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

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

    so b_i and w_i are case-wise slopes and weights, respectively;
  - **pairwise slopes**:

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

    so b_{ik} and w_{ik} are pairwise slopes and weights, respectively.

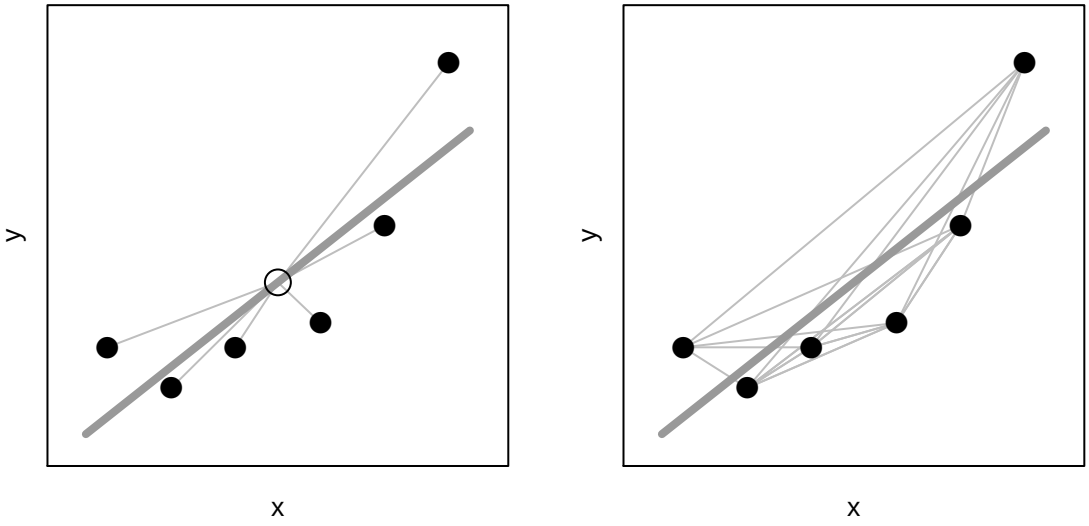

FIG 5. *Case-wise and pairwise average weighted slopes illustrated: Both plots show the same six points ("cases") as well as the OLS line fitted to them (fat gray). The left hand plot shows the case-wise slopes from the mean point (open circle) to the six cases, while the right hand plot shows the pairwise slopes between all 15 pairs.*

See Figure 5 for an illustration for samples. 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 the regressors.

In the LA homeless data, we can interpret the slope for the regressor `PercVacant`, say, in the following two ways:

(1) "Adjusted for all other regressors, the mean deviation of `Homeless` in relation to the mean deviation of `PercVacant` is estimated to be on 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 be on 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 assumption-laden 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."

Note on literature: The above formulas were used and modified to produce alternative slope estimates 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). Earlier, Wu (1986) used generalizations based on tuples rather than pairs of $(\vec{\boldsymbol{x}}_i', y_i)$ rows for the analysis of jackknife and bootstrap procedures (see his Section 3, Theorem 1). The formulas have a history in which Stigler (2001) includes Edgeworth, while Berman (1988) traces it back to a 1841 article by Jacobi written in Latin.

9. ASYMPTOTIC VARIANCES — PROPER AND IMPROPER

The following prepares the ground for an asymptotic comparison of assumption-laden with assumption-lean standard errors, one one regressor at a time.

9.1 Preliminaries: Adjustment Formulas for EOs and Their CLTs

The vectorized formulas for estimation offsets (23) can be written componentwise using adjustment as follows:

$$
\begin{aligned}
&\textit{Total EO}: && \hat\beta_j - \beta_j &&= \frac{\langle \boldsymbol{X}_{j\hat\bullet}, \boldsymbol{\delta}\rangle}{\|\boldsymbol{X}_{j\hat\bullet}\|^2},\\
&\textit{Noise EO}: && \hat\beta_j - \beta_j(\boldsymbol{X}) &&= \frac{\langle \boldsymbol{X}_{j\hat\bullet}, \boldsymbol{\epsilon}\rangle}{\|\boldsymbol{X}_{j\hat\bullet}\|^2},\\
&\textit{Model Deviation EO}: && \beta_j(\boldsymbol{X}) - \beta_j &&= \frac{\langle \boldsymbol{X}_{j\hat\bullet}, \boldsymbol{\eta}\rangle}{\|\boldsymbol{X}_{j\hat\bullet}\|^2}.
\end{aligned}
\tag{37}
$$

To see these identities directly, note the following, in addition to (36): $\boldsymbol{E}[\hat\beta_j|\boldsymbol{X}] = \langle \boldsymbol{\mu}, \boldsymbol{X}_{j\hat\bullet}\rangle/\|\boldsymbol{X}_{j\hat\bullet}\|^2$ and $\beta_j = \langle \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}_{j\hat\bullet}\rangle/\|\boldsymbol{X}_{j\hat\bullet}\|^2$, the latter due to $\langle \boldsymbol{X}_{j\hat\bullet}, \boldsymbol{X}_k\rangle = \delta_{jk}\|\boldsymbol{X}_{j\hat\bullet}\|^2$. Finally use $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$.

From (37), asymptotic normality of the coefficient-specific EOs can be separately expressed using population adjustment:

Corollary 9.1:

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

The equalities on the right side in the first and second case are based on (28). The first one will be needed for plug-in estimation. The sandwich form for matrices has been reduced to a ratio where the numerator corresponds to the "meat" and the squared denominator to the "breads".

9.2 Proper Asymptotic Variances in Terms of Adjusted Regressors

The CLTs of Corollary 9.1 contain three asymptotic variances of the same form with arguments $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$. This suggests using generic notation:

Definition 9.2: *Proper Asymptotic Variance and its Components.*

$$
\boxed{\boldsymbol{AV}_{lean}[\hat\beta_j; f^2] \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}})X_{j\bullet}^2]}{\boldsymbol{E}[\,X_{j\bullet}^2]^2},} \quad \textit{where } f^2(\vec{\boldsymbol{x}}) = m^2(\vec{\boldsymbol{x}}),\ \sigma^2(\vec{\boldsymbol{x}}) \textit{ or } \eta^2(\vec{\boldsymbol{x}}).
$$

Lemma 9.2: $\boldsymbol{AV}_{lean}[\hat{\beta}_j; m^2] \;=\; \boldsymbol{AV}_{lean}[\hat{\beta}_j; \sigma^2] \;+\; \boldsymbol{AV}_{lean}[\hat{\beta}_j; \eta^2]$.

9.3 Improper Asymptotic Variances in Terms of Adjusted Regressors

The goal is to provide an asymptotic analog for the "usual" standard error estimate of linear models theory, but in the assumption-lean framework. It derives from an estimate $\hat{\sigma}^2$ of the noise variance, $\hat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(N-p-1)$, which has the following limit for fixed p:

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

Squared standard error estimates are, in matrix and adjustment form, as follows:

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \hat{\sigma}^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}, \qquad \hat{\boldsymbol{SE}}^2_{lin}[\,\hat{\beta}_j] \;=\; \frac{\hat{\sigma}^2}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}. \tag{38}$$

Their scaled limits under lean 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 **improper** asymptotic variances because they provide valid standard errors only if the first and second order assumptions of linear models theory hold. Here is again a generic definition with an associated decomposition:

Definition 9.3: *Improper Asymptotic Variance and its Components.*

$$\boxed{\boldsymbol{AV}_{lin}[\hat{\beta}_j; f^2] \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X^2_{j\bullet}]},} \quad \textit{where } f^2(\vec{\boldsymbol{x}}) = m^2(\vec{\boldsymbol{x}}),\ \sigma^2(\vec{\boldsymbol{x}}) \textit{ or } \eta^2(\vec{\boldsymbol{x}}).$$

Lemma 9.3: $\boldsymbol{AV}_{lin}[\hat{\beta}_j; m^2] \;=\; \boldsymbol{AV}_{lin}[\hat{\beta}_j; \sigma^2] \;+\; \boldsymbol{AV}_{lin}[\hat{\beta}_j; \eta^2]$.

9.4 *RAV*: Comparison of Proper and Improper Asymptotic Variances

To examine the discrepancies between proper and improper asymptotic variances we form their ratios separately for each of the versions corresponding to $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$, hence we use again a generic form of the ratio:

Definition 9.4: *Ratio of Asymptotic Variances, Proper/Improper.*
For $f^2(\vec{\boldsymbol{x}}) = m^2(\vec{\boldsymbol{x}})$, $\sigma^2(\vec{\boldsymbol{x}})$ *or* $\eta^2(\vec{\boldsymbol{x}})$, *let*

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

Lemma 9.4: ***RAV*** *Decomposition.*

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

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

The three ***RAV*** terms can be interpreted as inner products between the three random variables

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

These are *not* correlations, and they are not upper bounded by +1; their natural bounds are rather 0 and ∞ (Section 9.6). A simplification is achieved by conditioning the left hand terms on $X_{j\bullet}^2$:

Definition and Lemma: *Let* $f_j^2(X_{j\bullet}^2) := \boldsymbol{E}[f^2(\vec{\boldsymbol{X}}) \,|\, X_{j\bullet}^2]$. *Then:*

$$m_j^2(X_{j\bullet}^2) = \eta_j^2(X_{j\bullet}^2) + \sigma_j^2(X_{j\bullet}^2) \quad \textit{and} \quad \boldsymbol{RAV}[\hat{\beta}_j, f^2] = \boldsymbol{RAV}[\hat{\beta}_j, f_j^2]. \tag{39}$$

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

9.5 The Meaning of *RAV*

The ratio $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$ shows by what multiple the improper asymptotic variance deviates from the proper one:

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

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

If, however, $\boldsymbol{RAV}[\hat{\beta}_j, m^2] = 1$, it does not follow that the model is well-specified. Well-specification to first and second order is sufficient but not necessary for asymptotic validity of the usual standard error. In particular, in view of Section 3.3, the following holds:

Lemma 9.5: *If* δ *and* $X_{j\bullet}$ *are independent, then* $\boldsymbol{RAV}[\hat{\beta}_j, m^2] = 1$.

9.6 The Range of *RAV*

The goal is to describe the extremes of the ***RAV***. These can be interpreted as extremes over scenarios of $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$, $\eta^2(\vec{\boldsymbol{X}})$, or, by (39), of $m_j^2(X_{j\bullet}^2)$, $\sigma_j^2(X_{j\bullet}^2)$, $\eta_j^2(X_{j\bullet}^2)$. The proposition below is stated for m_j^2:

Proposition 9.6: *If* $\boldsymbol{E}[X_{j\bullet}^2] < \infty$, *then*

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

(See Appendix D.2 for a proof and some technical subtleties.)

Corollary 9.6: *If* $\boldsymbol{E}[X_{j\bullet}^2] < \infty$ *and* $X_{j\bullet}$ *has unbounded support, then*

$$\sup_{m_j^2} \boldsymbol{RAV}[\hat{\beta}_j, m_j^2] \;=\; \infty.$$

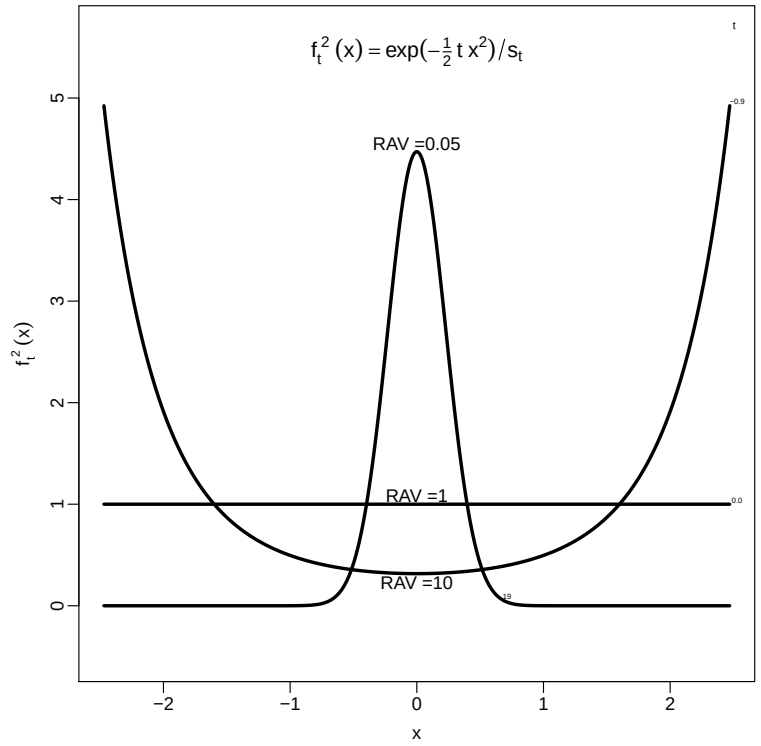

FIG 6. *A family of functions $f_t^2(x)$ that can be interpreted as heteroskedasticities $\sigma_j^2(X_{j\bullet})$, squared nonlinearities $\eta_j^2(X_{j\bullet})$, or conditional MSEs $m_j^2(X_{j\bullet})$: The family interpolates* ***RAV*** *from* 0 *to* ∞ *for* $x = X_{j\bullet} \sim N(0,1)$. *The three solid black curves show $f_t^2(x)$ that result in RAV=0.05, 1, and 10. (See Appendix D.3 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.*

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

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

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

What shapes of $m_j^2(X_{j\bullet})$ approximate these extremes? An intuitive answer can be guessed from Figure 6 for normally distributed $X_{j\bullet}$ to illustrate the corollary: 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.2 bears this out. The main concern is with usual standard errors that are optimistic, ***RAV*** > 1. The proposition shows that $X_{j\bullet}$-distributions with bounded support enjoy some protection from the worst case:

- 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 $\sqrt{3} \approx 1.732$.

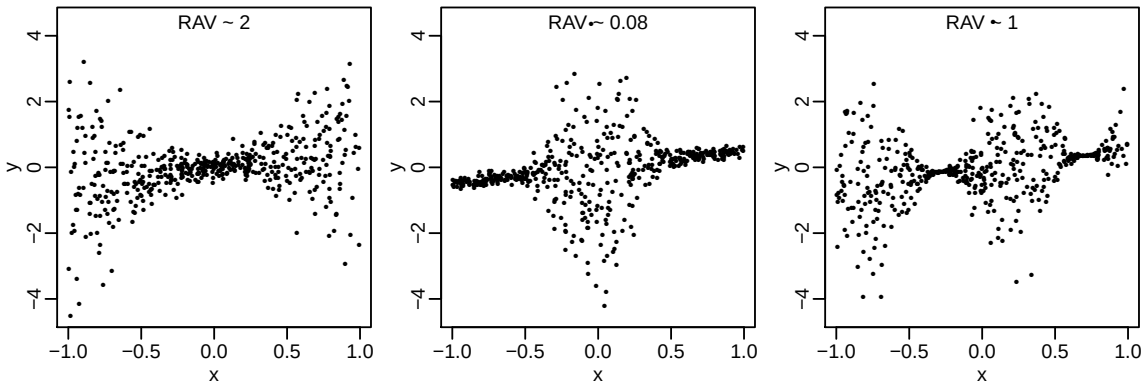

FIG 7. *The effect of heteroskedasticity on the sampling variability of slope estimates: How does the treatment of the heteroskedasticities as homoskedastic affect statistical inference?*
Left: High noise variance in the tails of the regressor distribution elevates the true sampling variability of the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

- 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 $\pm c$, 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 $X_{j\bullet}^2$.

9.7 Illustration of Factors that Drive the *RAV*

So far the results and illustrations for the $\boldsymbol{RAV}$ have been in terms of extreme scenarios for $m_j^2(X_{j\bullet}^2)$, which could also be interpreted as scenarios for $\sigma_j^2(X_{j\bullet}^2)$ and $\eta_j^2(X_{j\bullet}^2)$. This section illustrates the $\boldsymbol{RAV}$ in terms of potential data situations: 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 $\boldsymbol{RAV}$. According to the $\boldsymbol{RAV}$ decomposition of Lemma 9.4, $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$ is a mixture of $\boldsymbol{RAV}[\hat{\beta}_j, \sigma^2]$ and $\boldsymbol{RAV}[\hat{\beta}_j, \eta^2]$. Therefore:

- Heteroskedasticities with large $\sigma_j^2(X_{j\bullet}^2)$ in the tails of $X_{j\bullet}$ produce an upward contribution to $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$; heteroskedasticities with large $\sigma_j^2(X_{j\bullet}^2)$ near $X_{j\bullet}^2 = 0$ imply a downward contribution to $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$.
- 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 $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$; nonlinearities with large $\eta_j^2(X_{j\bullet}^2)$

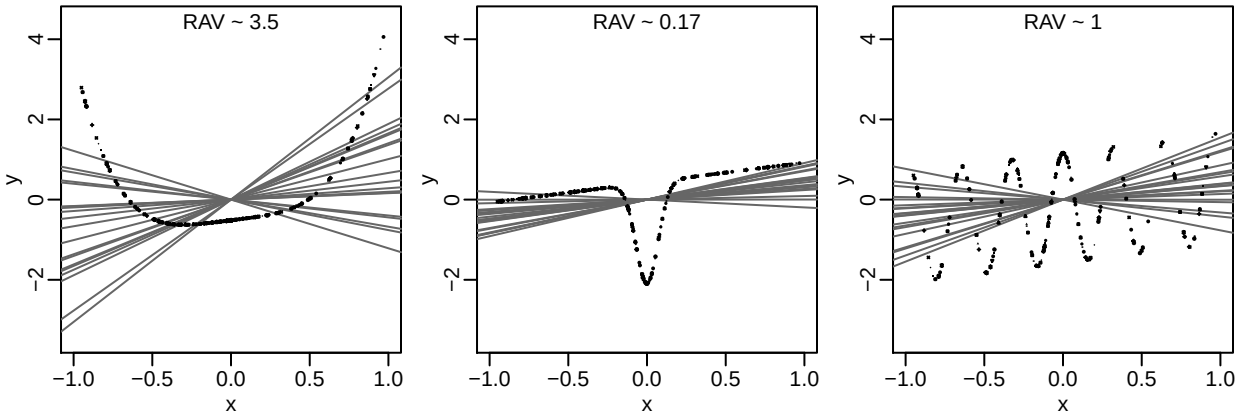

FIG 8. *The effect of nonlinearities on the sampling variability of slope estimates: The three plots show three different noise-free nonlinearities; each plot shows for one nonlinearity 20 overplotted datasets of size $N = 10$ and their fitted lines through the origin. The question is how the misinterpretation 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 (**RAV**$[\hat{\beta}_j, \eta^2] > 1$). Center: Strong nonlinearity near the center of the regressor distribution lowers the true sampling variability of the slope estimate below the usual standard error (**RAV**$[\hat{\beta}_j, \eta^2] < 1$). Right: An oscillating nonlinearity mimics homoskedastic random error to make the usual standard error coincidentally correct (**RAV**$[\hat{\beta}_j, \eta^2] = 1$).*

concentrated near $X_{j\bullet}^2 = 0$ imply a downward contribution to $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$.

These facts also suggest that, in practice, 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. This seems particularly natural for nonlinearities which in the simplest cases will be convex or concave. In addition it follows from the ***RAV*** decomposition of Lemma 9.4 that for fixed relative contributions $w_\sigma > 0$ and $w_\eta > 0$ either of $\boldsymbol{RAV}[\hat{\beta}_j, \sigma^2]$ or $\boldsymbol{RAV}[\hat{\beta}_j, \eta^2]$ is able to single-handedly pull $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$ to $+\infty$, whereas both have to be close to zero to pull $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$ toward zero. These considerations are heuristics for the observation that in practice $\widehat{\boldsymbol{SE}}_{lin}$ is more often too small than too large compared to $\widehat{\boldsymbol{SE}}_{sand}$.

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

The goal here is to write the ***RAV*** in adjustment form and estimate it with plug-in for use as a test statistic to decide whether the usual standard error is adequate. We will obtain one test per regressor.

The 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 information 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.

10.1 Sandwich Estimators in Adjustment Form and the $R\hat{A}V_j$ Test Statistic

To begin with, the adjustment versions of the asymptotic variances in the CLTs of Corollary 9.1 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 $\boldsymbol{X}_{j\hat{\bullet}}$, and rescaling by N:

$$\hat{\boldsymbol{SE}}_{sand}[\hat{\beta}_j]^2 \;=\; \frac{1}{N}\frac{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2]}{\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2]^{\,2}} \;=\; \frac{\langle(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^4}. \tag{40}$$

The squaring of N-vectors is meant to be coordinate-wise. Formula (40) is algebraically equivalent to (32).

The usual squared standard error estimate (38) is

$$\hat{\boldsymbol{SE}}_{lin}[\hat{\beta}_j]^2 \;=\; \frac{\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{(N\text{–}p\text{–}1)\,\|\boldsymbol{X}_{j\hat{\bullet}}\|^2} \;\sim\; \frac{1}{N}\frac{\hat{\boldsymbol{E}}[(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2]}{\hat{\boldsymbol{E}}[X_{j\hat{\bullet}}{}^2]} \;=\; \frac{\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{N\,\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,, \tag{41}$$

where the right hand forms are normalized to match (40), ignoring p. Thus the natural plug-in estimate of $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$ is

$$\boldsymbol{R\hat{A}V}_j \;:=\; \frac{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2\,]}{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2\,]\,\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2\,]} \;=\; N\,\frac{\langle(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\,\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,. \tag{42}$$

This is the proposed test statistic. Analogous to the population-level $\boldsymbol{RAV}[\hat{\beta}_j, m^2]$, the sample-level $\boldsymbol{R\hat{A}V}_j$ responds to associations between squared residuals and squared adjusted predictors, which parallels White's (1982, p. 12) remark that in linear regression his global misspecification test "is sensitive to forms of heteroskedasticity or model misspecification which result in correlations between the squared regression errors and the second order cross-products of the regressors."

10.2 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 $\boldsymbol{RAV}[\hat{\beta}_j, m^2] = 1$ by Section 3.3:

Proposition 10.2: *Under the null hypothesis H_0 that the population residuals δ and the adjusted regressor $X_{j\bullet}$ are independent, it holds:*

$$N^{1/2}\,(\boldsymbol{R\hat{A}V}_j - 1) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0,\, \frac{\boldsymbol{E}[\,\delta^4]}{\boldsymbol{E}[\,\delta^2]^2}\,\frac{\boldsymbol{E}[X_{j\bullet}{}^4]}{\boldsymbol{E}[X_{j\bullet}^2]^2} \;-\; 1)\right). \tag{43}$$

As always we ignore technical assumptions. A proof outline is in Appendix D.4.

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

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{sand}$	$\boldsymbol{R\hat{A}V}_j$	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 $\boldsymbol{R\hat{A}V}_j$ (10,000 permutations).

1. The larger the kurtosis of δ and/or $X_{j\bullet}$, more the asymptotic variance gets inflated, and hence the less likely is detection of first and second order model misspecification that resulting in standard error discrepancies.
2. Because 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 $X_{j\bullet}$, in which case both have symmetric two-point distributions (as both are centered). For such $X_{j\bullet}$ it follows that $\boldsymbol{RAV}[\hat{\beta}_j, m^2] = 1$ by Section 9.6.
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 otherwise diagnosed and tested.

10.3 An Approximate Permutation Distribution of the *RAV* Test Statistic

The asymptotic result of Proposition 10.2 provides qualitative insights, but it is not suitable for practical application because the null distribution of $\boldsymbol{R\hat{A}V}_j$ 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 $\boldsymbol{X}_{j\hat{\bullet}}{}^2$ are formed, a draw from the conditional null distribution of $\boldsymbol{R\hat{A}V}_j$ is obtained by randomly permuting one of the vectors and forming the inner product with the other vector, rescaled by a factor $N/(\|\boldsymbol{r}\|^2\|\boldsymbol{X}_{j\hat{\bullet}}\|^2)$. 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. Additionally, the permutation distribution yields an easy diagnostic of non-normality (see Appendix E for examples).

Table 4 shows the results for the LA Homeless data. Values of $\boldsymbol{R\hat{A}V}_j$ that fall outside the middle 95% range of their permutation null distributions are marked with asterisks. Surprisingly, the values of approximately 2 for the $\boldsymbol{R\hat{A}V}_j$ of `PercVacant` and `PercCommercial` are not statistically significant.

10.4 Generalizations of *RAV* Tests

The ***RAV*** test proposed here seems to be novel. It is not a special case of White's (1980b) global heteroskedasticity test, nor of his misspecification test for general ML estimation (White 1982). The latter uses a test statistic based on the

sum of the Hessian and outer product forms of the information matrix, which works on the matrix-inverse scale of asymptotic variances and is hence incapable of comparing specific "usual" (model-trusting) and "proper" (assumption-lean) asymptotic variances of specific parameters. Generalized ***RAV*** tests are conceivable for general MM estimators by forming ratios $\hat{\boldsymbol{AV}}_{jj}/(\hat{\boldsymbol{\Lambda}}^{-1})_{jj}$ using notation of Sections 5.4 (29) and 6.2 (34). We do not have results for ***RAV*** tests in this generality, however.

11. ISSUES WITH ASSUMPTION-LEAN STANDARD ERRORS

Model-robustness is a highly desirable property of the sandwich estimator, but as always there is no free lunch. As Kauermann and Carroll (2001) have shown, a cost of the sandwich estimator may be **inefficiency when the assumed model is correct**. Using plug-in in asymptotic variances can lead to standard errors that are too small/optimistic because the variability from plug-in is not accounted for. Sandwich estimators should therefore be accurate only when the sample size is sufficiently large. This fact suggests that use of the model-trusting standard error should be kept in mind if there is evidence in its favor, for example, through the ***RAV*** test of Section 10. [Kauermann and Carroll's analysis is for fixed regressors and treats heteroskedasticity only, but its message is valid because it speaks to performance under well-specification.]

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 outlying observations and heavy-tailed error distributions: The statistic $\hat{\boldsymbol{SE}}^2_{sand}[\hat{\beta}_j]$ (40) is a ratio of fourth order quantities of the data, whereas $\hat{\boldsymbol{SE}}^2_{lin}[\hat{\beta}_j]$ (41) is "only" a ratio of second order quantities. [Note we are here concerned not with non-robustness of parameter estimates but their standard error estimates.] It appears, therefore, that the two types of robustness are in conflict: Model-robust standard error estimators are highly non-robust compared to their model-trusting analogs. This is a large issue which we can only raise but not solve in this space. Here are a few observations and suggestions:

- If model-robust standard errors are not classically robust, we find anecdotally that the converse may hold also: 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 and its model-robust version obtained from the x-y-bootstrap. (We used the function `rlm` in the **R** *Language* (2008)).
- Yet classical robust regression may confer partial robustness to the sandwich standard error because it limits the size of residuals by capping them with a bounded ψ function. This addresses robustness to outlyingness in the vertical (y) direction.
- Robustness to outlyingness in the horizontal ($\vec{\boldsymbol{x}}$) direction could be achieved by using bounded-influence regression (see, e.g., Krasker and Welsch 1982, and references therein) which automatically downweights observations in high-leverage positions, or by using some other downweighting scheme to control the effects of high-leverage points.
- Robustness to horizontal outlyingness could also be addressed by transform-

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	t_{lin}	t_{boot}	t_{sand}
(Intercept)	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; regressors are transformed with cdfs.

ing the regressor variables to bounded ranges. Taking a cue from Proposition 9.6, one might search for transformations that obviate the need for an assumption-lean standard error in the first place.

As an illustration of the last point, we transformed the regressors of the LA Homeless data with their empirical cdfs to achieve approximately uniform marginal distributions up to discreteness. The transformed data are no longer i.i.d., but the point is to show the potential 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. (Note that bounded ranges are really needed for the adjusted regressors, but transformation of the raw regressors is likely to achieve this when the collinearities are not extreme.)

12. SUMMARY AND OUTLOOK

For linear OLS, the sandwich estimator of standard error is widely known to be heteroskedasticity-consistent, but it is less known to be also nonlinearity-consistent. Nonlinearity is the more severe issue which calls into question the meaning of slopes of linear fits and which invalidates regressor ancillarity. As a consequence, linear slopes require a new interpretation, slopes depend on the regressor distribution, conditioning on the regressors is no longer justified, nonlinearity generates a contribution to sampling variability that is unrelated to the conditional distribution of the response given the regressors, and the "usual" model-trusting standard error may be asymptotically incorrect. Thus the idea that models are approximations and may generally be misspecified to some degree may suggest resorting to model-robust standard errors of the sandwich variety.

These facts generalize to arbitrary MM estimation, including ML, quasi-likelihood and instrumental variable regression. The notion is that a set of moment conditions is well-specified for a joint regressor-response distribution $\boldsymbol{P}$ if the "design-conditional parameter" $\boldsymbol{\theta}(\boldsymbol{X})$ is the same irrespective of designs $\boldsymbol{X}$, in which case it agrees with the population parameter $\boldsymbol{\theta}(\boldsymbol{P})$. In case the moment condition is misspecified in this sense, the conditional parameter $\boldsymbol{\theta}(\boldsymbol{X})$ is not constant and hence has genuine sampling variability stemming from the marginal regressor distribution rather than the conditional response distribution.

For OLS it is possible to identify the nature of misspecifications that render standard errors too optimistic or too pessimistic, or neither. In the latter case the misspecification does not affect the validity of the usual standard error. This

possibility suggests that general-purpose model diagnostics are not the best route for establishing the validity of inference based on the usual standard error. Rather, a specific test is needed, such as the ***RAV*** test proposed here.

Since White's seminal work, research into misspecification has progressed far and in many forms by addressing specific classes of model deviation: 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 model deviations 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 arises a need for assumption-lean/model-robust inference. Even when complex modeling is possible, simple questions sometimes call for simple models, in which case again one may want to look for assumption-lean inference.

There exist, finally, areas of statistics research where assumption-laden theory appears frequently:

- Bayes inference, when it relies on uninformative priors, is asymptotically equivalent to assumption-laden 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. Interesting developments are taking place: 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 robust statistics by, for example, El Karoui et al. (2013), Donoho and Montanari (2014), and Loh (2015).

Thus there remains work to be done especially in some of today's most lively research areas. Even within the narrower, non-Baysian 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 satisfactorily realized.

Acknowledgments: We are grateful to Gemma Moran and Bikram Karmakar for their help in the generalizations of Section 5.

REFERENCES

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

[2] Berk, R. A. 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.)

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

[4] Bickel, P. J. and Götze, F. and van Zwet, W. R. (1997). *Statistica Sinica* **7**, 1–31.

[5] 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.

[6] Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. **arXiv:1503.06426**

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

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

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

[10] 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.

[11] De Blasi, P. (2013). Discussion on article "Bayesian inference with misspecified models" by Stephen G. Walker. *Journal of Statistical Planning and Inference* **143**, 1634–1637.

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

[13] 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.

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

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

[16] 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.

[17] Harrison, X. and Rubinfeld, X. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management* **5**, 81–102.

[18] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

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

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

[21] 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.

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

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

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

[25] Hoff, P. and Wakefield, J. (2013). Bayesian sandwich posteriors for pseudo-true parameters — A discussion of "Bayesian inference with misspecified models" by Stephen Walker. *Journal of Statistical Planning and Inference* **143**, 1638–1642.

[26] 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.

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

[28] 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.

[29] Kent, J. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69** (1), 19–27.

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

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

[32] Loh, P. (2015). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. **arXiv:1501.00312**

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

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

[35] MacKinnon, J. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29**, 305–325.

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

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

[38] 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.

[39] 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.

[40] 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.

[41] Stigler, S. M. (2001). Ancillary History. In *State of the Art in Probability and Statistics: Festschrift for Willem R. van Zwet* (M. DeGunst, C. Klaassen and A. van der Vaart, eds.), 555–567.

[42] Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference* **143**, 1621–1633.

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

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

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

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

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

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

[49] 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.

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

illustrates the ***RAV*** test for the Boston Housing data. Values of $\boldsymbol{R}\hat{\boldsymbol{A}}\boldsymbol{V}_j$ that fall outside the middle 95% range of their permutation null distributions are marked with asterisks.

Table 8 illustrates the ***RAV*** test for the Boston Housing data. Values of $\boldsymbol{R}\hat{\boldsymbol{A}}\boldsymbol{V}_j$ 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 $\boldsymbol{X}$ 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:

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	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.

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{sand}$	$\boldsymbol{R\hat{A}V}_j$	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).

(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:

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

where $\boldsymbol{X}$ acts as the ancillary $\boldsymbol{A}$ and $p_{\boldsymbol{X}}$ 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

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

and let

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

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 $\boldsymbol{X}_{j\hat{\bullet}}$ by collecting all regressor columns other than $\boldsymbol{X}_j$ in a $N \times p$ random regressor matrix

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

and let

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

(Note the use of hat notation "$\hat{\bullet}$" to distinguish it from population-based adjustment "$\bullet$".) The response vector $\boldsymbol{Y}$ can be sample-adjusted similarly, and we may denote it by $\boldsymbol{Y}_{\hat{\bullet}_{-j}}$ to indicate that $\boldsymbol{X}_j$ 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: $\boldsymbol{E}[Y|\vec{\boldsymbol{X}}] = \mu(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} \mu_o(\vec{\boldsymbol{X}})$. In this class the following holds:*

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

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: $\boldsymbol{E}[Y|\vec{\boldsymbol{X}}] = \mu(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} \mu_o(\vec{\boldsymbol{X}})$. In this class the following holds:

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

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 Proofs of *RAV*-Range Propositions in Section 9.6

The ***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 9.6 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 $\boldsymbol{P}$-min and $\boldsymbol{P}$-max of $X_{j\bullet}^2$ do not carry positive probability mass. This is a largely technical condition because even for discrete predictors X_j the adjusted squared version $X_{j\bullet}^2$ will have a continuous distribution if there exists just one other predictor that is continuous and non-orthogonal (partly collinear) to X_j.

Lemma D.2.1: *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:

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

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

Lemma D.2.2:

(a) Define a one-parameter family

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

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

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

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:

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

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

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

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

Proof of part (b): Preliminary observations:

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

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

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

D.3 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 $t > -1$:

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

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

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

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.4 Proof of Asymptotic Normality of $R\hat{A}V_j$, Section 10.2

We will need notation for each observation's population-adjusted regressors: $\boldsymbol{X}_{j\bullet} = (X_{1,j\bullet}, ..., X_{N,j\bullet})' = \boldsymbol{X}_j - \boldsymbol{X}_{\text{-}j}\boldsymbol{\beta}_{\text{-}j\bullet}$. The following distinction is elementary but important: The component variables of $\boldsymbol{X}_{j\bullet} = (X_{i,j\bullet})_{i=1...N}$ are i.i.d. as they are population-adjusted, whereas the component variables of $\boldsymbol{X}_{j\hat{\bullet}} = (X_{i,j\hat{\bullet}})_{i=1...N}$ 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:

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

We recall (42) for reference in the following form:

$$\boldsymbol{R\hat{A}V}_j \;=\; \frac{\frac{1}{N}\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\frac{1}{N}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 \; \frac{1}{N}\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,. \tag{44}$$

For the denominators it is easy to show that

$$\begin{aligned} \tfrac{1}{N}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 &\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[\,\delta^2\,], \\ \tfrac{1}{N}\|\boldsymbol{X}_{j\hat{\bullet}}\|^2 &\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[\,X_{j\bullet}^2\,]. \end{aligned} \tag{45}$$

For the numerator a CLT holds based on

$$\tfrac{1}{N^{1/2}}\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle \;=\; \tfrac{1}{N^{1/2}}\langle (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^2, \boldsymbol{X}_{j\bullet}{}^2\rangle + O_P(N^{-1/2}). \tag{46}$$

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

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

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 i.i.d. sampled pairs, hence $\{(\delta_i^2, X_{i,j\bullet}{}^2)\}_{i=1...N}$ are N i.i.d. sampled pairs as well. Using the denominator terms (45) 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 (46), using notation of Sections C.1 and C.2, in particular $\boldsymbol{X}_{j\bullet} = \boldsymbol{X}_j - \boldsymbol{X}_{\text{-}j}\boldsymbol{\beta}_{\text{-}j\bullet}$ and $\boldsymbol{X}_{j\hat{\bullet}} = \boldsymbol{X}_j - \boldsymbol{X}_{\text{-}j}\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}}$:

$$\begin{aligned} \langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle &= \langle\, ((\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2,\, (\boldsymbol{X}_{j\bullet} - \boldsymbol{X}_{\text{-}j}(\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \boldsymbol{\beta}_{\text{-}j\bullet}))^2\,\rangle \\ &= \langle\, \boldsymbol{\delta}^2 + (\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2 - 2\,\boldsymbol{\delta}\,(\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})), \\ &\qquad \boldsymbol{X}_{j\bullet}{}^2 + (\boldsymbol{X}_{\text{-}j}(\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \hat{\boldsymbol{\beta}}_{\text{-}j\bullet}))^2 - 2\,\boldsymbol{X}_{j\bullet}(\boldsymbol{X}_{\text{-}j}(\hat{\boldsymbol{\beta}}_{\text{-}j\hat{\bullet}} - \boldsymbol{\beta}_{\text{-}j\bullet}))\,\rangle \\ &= \langle\, \boldsymbol{\delta}^2, \boldsymbol{X}_{j\bullet}{}^2\,\rangle + ... \end{aligned} \tag{47}$$

Among the 8 terms in "...", each contains at least one subterm of the form $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ or $\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} - \boldsymbol{\beta}_{-j\bullet}$, 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 $N^{1/2}$:

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

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 (47) 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:

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

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

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

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

APPENDIX E: NON-NORMALITY OF CONDITIONAL NULL DISTRIBUTIONS OF $R\hat{A}V_J$

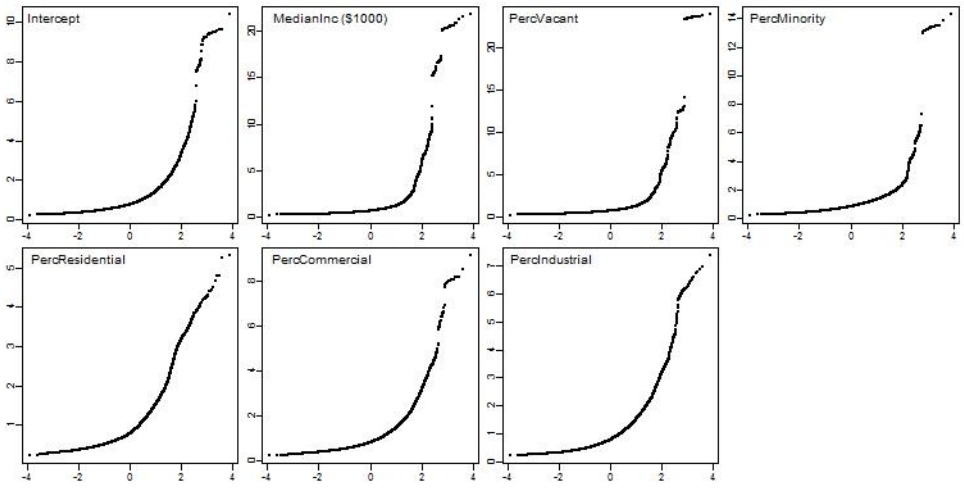

FIG 9. *Permutation distributions of* $\boldsymbol{R\hat{A}V_j}$ *for the LA Homeless Data*

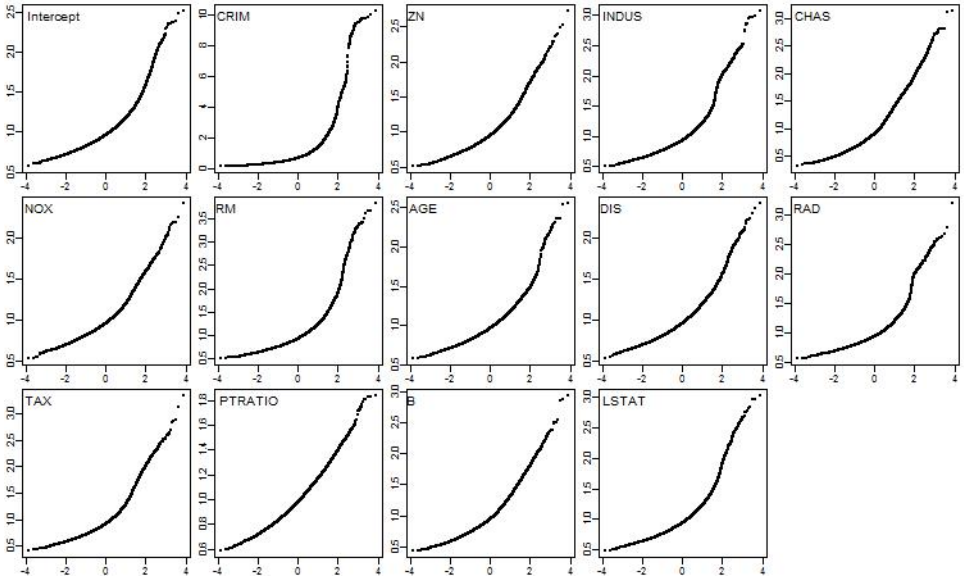

FIG 10. *Permutation distributions of* $\boldsymbol{R\hat{A}V_j}$ *for the Boston Housing Data*