Models as Approximations — A Conspiracy of Random Predictors and Model Violations 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.

We review and interpret the early insights of Halbert White who over thirty years ago inaugurated a form of statistical inference for regression models that is asymptotically correct even under "model misspecification," that is, under the assumption that models are approximations rather than generative truths. This form of inference, which is pervasive in econometrics, relies on the "sandwich estimator" of standard error. Whereas linear models theory in statistics assumes models to be true and predictors to be fixed, White's theory permits models to be approximate and predictors to be random. Careful reading of his work shows that the deepest consequences for statistical inference arise from a synergy — a "conspiracy" — of nonlinearity and randomness of the predictors which invalidates the ancillarity argument that justifies conditioning on the predictors when they are random. An asymptotic comparison of standard error estimates from linear models theory and White's asymptotic theory shows that discrepancies between them can be of arbitrary magnitude. In practice, when there exist discrepancies, linear models theory tends to be too liberal but occasionally it can be too conservative as well. A valid alternative to the sandwich estimator is provided by the "pairs bootstrap"; in fact, the sandwich estimator can be shown to be a limiting case of the pairs bootstrap. Finally we give

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.

meaning to regression slopes when the linear model is an approximation rather than a truth. — We limit ourselves to linear LS regression, but many qualitative insights hold for most forms of regression.

AMS 2000 subject classifications: Primary 62J05, 62J20, 62F40; secondary 62F35, 62A10.
Key words and phrases: Ancillarity of predictors, First and second order incorrect models, Model misspecification, Misspecification tests, Econometrics, Sandwich estimator of standard error, Pairs bootstrap.

1. INTRODUCTION

The classical Gaussian linear model reads as follows:

$$(1) \qquad \boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}_N, \sigma^2 \boldsymbol{I}_{N\times N}) \qquad (\boldsymbol{y}, \boldsymbol{\epsilon} \in \mathbb{R}^N,\ \boldsymbol{X} \in \mathbb{R}^{N\times(p+1)},\ \boldsymbol{\beta} \in \mathbb{R}^{p+1}).$$

Two aspects will be important: (1) the model is assumed correct, including linearity of the response means in the predictors and independence, homoskedasticity and Gaussianity of the errors; (2) the predictors are treated as known constants, even when they are as random as the response. Statisticians have long enjoyed the fruits that can be harvested from this model and they have taught it as fundamental at all levels of statistical education. Curiously little known to many statisticians is the fact that a different framework is adopted and a different statistical education is taking place in the parallel universe of econometrics. For over three decades, starting with Halbert White's (1980a,b;1981;1982) seminal articles, econometricians have used multiple linear regression without making the many assumptions of classical linear models theory. While statisticians use **assumption-laden exact finite sample inference**, econometricians use **assumption-lean asymptotic inference** based on the so-called "sandwich estimator" of standard error. In our experience most statisticians have heard of the sandwich estimator but do not know its purpose, use, and underlying theory. A first goal of the present exposition is therefore to convey an understanding of an assumption-lean framework in a language that is intelligible to statisticians. The approach is to interpret linear regression in a semi-parametric fashion as extraction of a parametric linear part of a general nonlinear response surface. The modeling assumptions can then be reduced to i.i.d. sampling from largely arbitrary joint $(\vec{\boldsymbol{X}}, Y)$ distributions that satisfy a few moment conditions. In this assumption-lean framework the sandwich estimator produces asymptotically correct standard errors.

A second goal of this exposition is to connect the assumption-lean framework to a form of statistical inference in linear models that is known to statisticians but appreciated by few: the "pairs bootstrap." As the name indicates, the pairs bootstrap consists of resampling pairs $(\vec{\boldsymbol{x}}_i, y_i)$, in contrast to the "residual bootstrap" which resamples residuals r_i. Among the two, the pairs bootstrap is the less promoted even though asymptotic theory exists to justify both under different assumptions (see, for example, Freedman 1981, Mammen 1993). It is intuitively clear that the pairs bootstrap can be asymptotically justified in the assumption-lean framework, and for this reason it produces standard error estimates that solve the same problem as the sandwich estimator. Indeed, we establish a connection that shows the sandwich estimator to be the asymptotic limit of the M-of-N pairs bootstrap when $M \to \infty$. We will use the general term "**assumption-lean estimator**" to refer to either the sandwich estimator or the pairs bootstrap estimator of standard error.

A third goal of this article is to theoretically and practically compare the assumption-lean estimators with the linear models estimator. We define a ratio of asymptotic variances — "***RAV***" for short — that describes the discrepancies between the two types of standard error estimators in the asymptotic limit. If there exists a discrepancy, $\boldsymbol{RAV} \neq 1$, it will be assumption-lean estimators (sandwich or pairs bootstrap) that are asymptotically correct, and the linear models estimator is then indeed asymptotically incorrect. If $\boldsymbol{RAV} \neq 1$, there exist deviations from the linear model in the form of nonlinearities and/or heteroskedasticities. If $\boldsymbol{RAV} = 1$, the linear models estimator is asymptotically correct, but this does not imply that the linear model is correct, the reason being that nonlinearities and heteroskedasticities can combine to produce a coincidentally correct size for the linear models estimator. Importantly, the ***RAV*** is specific to each regression coefficient because the magnitudes of the discrepancies can vary from coefficient to coefficient in the same model.

A fourth 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 classical standard error of a specific coefficient. While the result can be called a "misspecification test" in the tradition of econometrics, it is more usefully viewed as guidance to the better standard error. The sandwich estimator comes with a cost — vastly increased non-robustness — which makes it desirable to use the linear models standard error when possible. However, a procedure that chooses the type of standard error depending on the outcome of a pre-test raises new performance issues that require future research. (Simpler is Angrist and Pischke's proposal (2009) to choose the larger of the two standard errors; their procedure forfeits the possibility of detecting the classical standard error as too conservative. MacKinnon and White (1985) on the other hand recommend using a sandwich estimator even if no misspecification is detected.)

A fifth and final goal of this article is to propose answers to questions and objections that would be natural to statisticians who hold the following tenets:

1. Models need to be "correct" for inference to be meaningful. The implication is that assumption-lean approaches are misguided because inference in "misspecified models" are meaningless.
2. Predictors in regression models should or can be treated as fixed even if they are random. The implication is that inference which treats predictors as random is unprincipled or at best superfluous.

A strong proponent of the first tenet is the late David Freedman (2006). While his insistence on intellectual honesty and rigor is admirable, we will counterargue based on White (1980a,b; 1981; 1982) that inference under misspecification can be meaningful and rigorous. Abandoning the negative rhetoric of "misspecification", we adopt the following alternative view:

1. Models are always approximations, not truths (Box 1979; Cox 1995).
2. If models are approximations, regression slopes still have meaningful interpretations.
3. If models are approximations, it is prudent to use inference that is less dependent on model correctness.

In fact, neither the second nor the third point depends on the degree of approximation: meaningful interpretation and valid inference can be provided for regression slopes whether models are good or bad approximations. (This is of course not to say that every regression analysis is meaningful.)

The second tenet above, conditionality on predictors, has proponents to various degrees. More

forceful ones hold that conditioning on the predictors is a necessary consequence of the ancillarity principle; others hold that the principle confers license, not a mandate. The ancillarity principle says in simplified terms that valid inference results from conditioning on statistics whose distribution does not involve the parameters of interest. When the predictors in a regression are random, their distribution is ancillary for the regression parameters, hence conditioning on the predictors is necessary or at least permitted. This argument, however, fails when the parametric model is only an approximation and the predictors are random. It will be seen that under these circumstances the population slopes do depend on the predictor distribution which is hence not ancillary. This effect does not exist when the conditional mean of the response is a linear function of the predictors or the predictors are truly nonrandom.

This article continuous as follows: Section 2 illustrates discrepancies between standard error estimates with real data examples. Section 3 sets up the semi-parametric population framework in which LS approximation extracts a parametric linear component. Section 4 shows how nonlinear conditional expectations invalidate ancillarity of the predictor distribution. Section 5 derives a decomposition of asymptotic variance of LS estimates into contributions from noise and from nonlinearity, and central limit theorems associated with the decomposition. Section 6 introduces the simplest version of the sandwich estimator of standard error and shows how it is a limiting case of the M-of-N pairs bootstrap. Section 7 expresses parameters, estimates, asymptotic variances and CLTs in the language of predictor adjustment, which is critical in order to arrive at expressions that speak to individual predictors in a transparent fashion. In Section 8 the language of adjustment allows us to define the ***RAV***, that is, the ratio of proper (assumption-lean) and improper (linear models-based) asymptotic variances and to analyze conditions under which linear models theory yields standard errors that are too liberal (often) or too conservative (less often). Section 9 turns the ***RAV*** into a simple test statistic with an asymptotically normal null distribution under "well-specification". The penultimate Section 10 proposes an answer to the question of what the meaning of regression coefficients is when the linear model is an approximation rather than a generative truth. The final Section 11 has a brief summary and ends with a pointer to problematic aspects of the sandwich estimator. Throughout we use precise notation for clarity. The notation may give the impression of a technical article, but all technical results are elementary and all limit theorems are stated informally without regularity conditions.

2. DISCREPANCIES BETWEEN STANDARD ERRORS ILLUSTRATED

Table 1 shows regression results for a dataset in 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 pairs 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 predictors `PercVacant`, `PercCommercial` and `PercIndustrial`, and $\boldsymbol{SE}_{boot}, \boldsymbol{SE}_{sand} < \boldsymbol{SE}_{lin}$ for `Intercept`,

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	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
Comparison of Standard Errors for the LA Homeless Data.

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 predictors PercCommercial and PercIndustrial have commanding t-values of 2.700 and 2.818, respectively, which are reduced to unconvincing values below 1.9 and 1.6, respectively, if the pairs 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.

Table 2 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 predictors there is no major discrepancy among all three standard errors: ZN, NOX, B, and even for CRIM, $\boldsymbol{SE}_{lin}$ falls between the somewhat discrepant values of $\boldsymbol{SE}_{boot}$ and $\boldsymbol{SE}_{sand}$.

We conclude: (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 show substantial discrepancies; (3) the discrepancies are specific to predictors.

3. THE SEMI-PARAMETRIC POPULATION FRAMEWORK

3.1 Targets of Estimation

Before standard errors can be meaningfully compared it is necessary to describe a semi-parametric framework and define targets of estimation. The latter will no longer be parameters in a generative model but statistical functionals that are well-defined for a large nonparametric class of data distributions. A seminal work that inaugurated this approach is P.J. Huber's (1967) article whose title is worth citing in full: "The behavior of maximum likelihood estimation under nonstandard conditions." The "nonstandard conditions" are essentially arbitrary distributions for which certain moments exist.

In a semi-parametric population framework for linear regression with random predictors the ingredients are random variables X_1, ..., X_p and Y, where Y is singled out as the response. For now

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{boot}$	$\boldsymbol{SE}_{sand}$	$\frac{\boldsymbol{SE}_{boot}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{lin}}$	$\frac{\boldsymbol{SE}_{sand}}{\boldsymbol{SE}_{boot}}$	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 2

Comparison of Standard Errors for the Boston Housing Data.

the only assumption is that these variables 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 predictors have a full rank covariance matrix. We write

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

for the *column* random vector consisting of the predictor variables, with a constant 1 prepended to accommodate an intercept term. Values of the random vector $\vec{\boldsymbol{X}}$ will be denoted by lower case $\vec{\boldsymbol{x}} = (1, x_1, ..., x_p)^T$. We write any function $f(X_1, ..., X_p)$ of the predictors equivalently as $f(\vec{\boldsymbol{X}})$ as the prepended constant 1 is irrelevant. Correspondingly we also use the notations

$$\boldsymbol{P} = \boldsymbol{P}(\mathrm{d}y, \mathrm{d}\vec{\boldsymbol{x}}), \quad \boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}}), \quad \boldsymbol{P}(\mathrm{d}y \,|\, \vec{\boldsymbol{x}}) \qquad \text{and} \qquad \boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}, \quad \boldsymbol{P}_{\vec{\boldsymbol{X}}}, \quad \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}, \tag{2}$$

for 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. Nonsingularity of the predictor covariance matrix is equivalent to nonsingularity of the cross-moment matrix $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$.

Among functions of the predictors, an important one is the *best* $L_2(\boldsymbol{P})$ *approximation* to the response Y, which is the conditional expectation of Y given $\vec{\boldsymbol{X}}$:

$$\mu(\vec{\boldsymbol{X}}) \;:=\; \mathrm{argmin}_{f(\vec{\boldsymbol{X}})\in L_2(\boldsymbol{P})} \boldsymbol{E}[(Y - f(\vec{\boldsymbol{X}}))^2] \;=\; \boldsymbol{E}[\,Y \,|\, \vec{\boldsymbol{X}}\,]\,. \tag{3}$$

This is also called the "response surface." Importantly we do *not* assume that $\mu(\vec{\boldsymbol{X}})$ is linear in $\vec{\boldsymbol{X}}$.

Among linear functions $l(\vec{\boldsymbol{X}}) = \boldsymbol{\beta}^T\vec{\boldsymbol{X}}$, one stands out as the *best linear* $L_2(\boldsymbol{P})$ *approximation* or the *population LS linear approximation* to Y:

$$\boldsymbol{\beta}(\boldsymbol{P}) \;:=\; \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \boldsymbol{E}[(Y - \boldsymbol{\beta}^T\vec{\boldsymbol{X}})^2] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\boldsymbol{E}[\,\vec{\boldsymbol{X}}Y\,]\,. \tag{4}$$

The right most expression follows from the normal equations $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}Y] = \boldsymbol{0}$ that are the stationarity conditions for minimizing the population LS criterion $\boldsymbol{E}[(Y - \boldsymbol{\beta}^T\vec{\boldsymbol{X}})^2] = -2\boldsymbol{\beta}^T\boldsymbol{E}[\vec{\boldsymbol{X}}Y] + \boldsymbol{\beta}^T\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{\beta} + \mathrm{const}$.

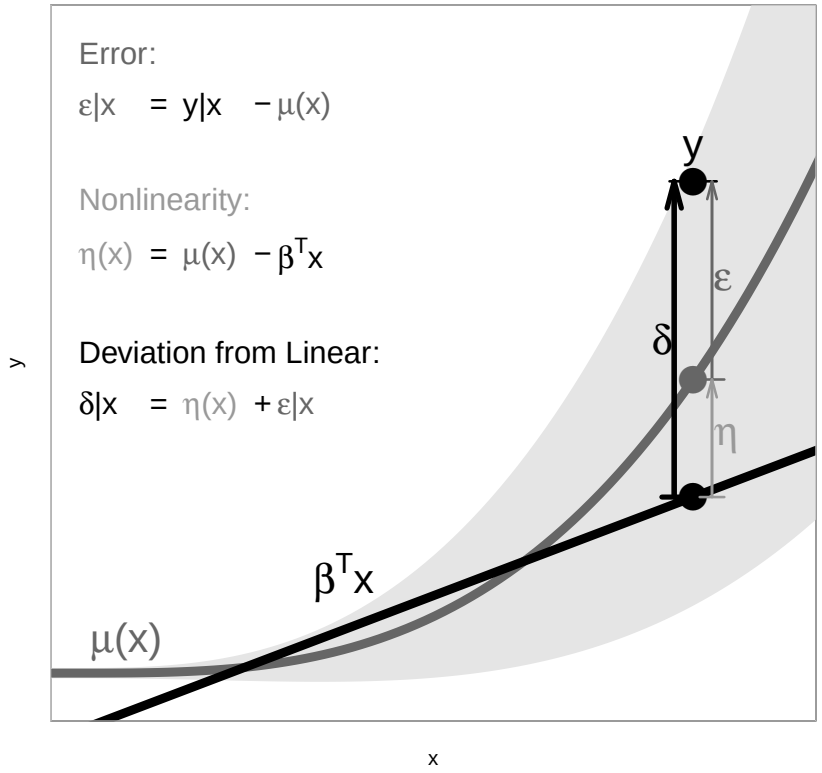

FIG 1. *Illustration of the decomposition* (5).

By abuse of terminology, we use the expressions "population coefficients" for $\boldsymbol{\beta}(\boldsymbol{P})$ and "population approximation" for $\boldsymbol{\beta}(\boldsymbol{P})^T\vec{\boldsymbol{X}}$, omitting the essential terms "linear" and "LS" or "$L_2(\boldsymbol{P})$" to avoid cumbersome language. We will often write $\boldsymbol{\beta}$, omitting the argument $\boldsymbol{P}$ when it is clear from the context that $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$.

The population coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ form a *statistical functional* that is defined for a large class of data distributions $\boldsymbol{P}$. Below it will be made obvious that $\boldsymbol{\beta}(\boldsymbol{P})$ is *not* the target of estimation in classical linear models theory.

3.2 The Noise-Nonlinearity Decomposition

The response Y has the following natural decompositions:

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

These equalities define the random variables $\eta = \eta(\vec{\boldsymbol{X}}) = \mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}^T\vec{\boldsymbol{X}}$, called ***nonlinearity***, $\epsilon = Y - \mu(\vec{\boldsymbol{X}})$, called ***noise***, and

$$
\delta = \epsilon + \eta, \tag{6}
$$

for which there does not exist a standard term, hence the name "total deviation" (from linearity) may suffice. An attempt to depict the decompositions (5) for a single predictor is given in Figure 1.

The noise ϵ is not assumed homoskedastic, and its conditional distributions $\boldsymbol{P}(d\epsilon|\vec{\boldsymbol{X}})$ can be quite arbitrary except for being centered and having second moments almost surely:

$$\boldsymbol{E}[\,\epsilon\,|\vec{\boldsymbol{X}}] \;\overset{\boldsymbol{P}}{=}\; 0, \tag{7}$$

$$\sigma^2(\vec{\boldsymbol{X}}) \;:=\; \boldsymbol{V}[\,\epsilon\,|\vec{\boldsymbol{X}}] \;=\; \boldsymbol{E}[\,\epsilon^2\,|\,\vec{\boldsymbol{X}}] \;\overset{\boldsymbol{P}}{<}\; \infty. \tag{8}$$

In addition to the ***conditional variance*** $\sigma^2(\vec{\boldsymbol{X}})$, we will need the ***conditional mean squared error*** $m^2(\vec{\boldsymbol{X}})$ of the population LS linear function, and its variance-bias^2 decomposition associated with (6):

$$m^2(\vec{\boldsymbol{X}}) \;:=\; \boldsymbol{E}[\,\delta^2\,|\,\vec{\boldsymbol{X}}] \;=\; \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}}). \tag{9}$$

3.3 Interpretations and Properties of the Decompositions

Equations (5) above can be given the following ***semi-parametric interpretation***:

$$\underbrace{\mu(\vec{\boldsymbol{X}})}_{\textit{semi-parametric part}} \;=\; \underbrace{\boldsymbol{\beta}^T\vec{\boldsymbol{X}}}_{\textit{parametric part}} \;+\; \underbrace{\eta(\vec{\boldsymbol{X}})}_{\textit{nonparametric part}}. \tag{10}$$

Thus the proposed purpose of linear regression is to extract the parametric part of the response surface and provide statistical inference for the parameters in the presence of a nonparametric part.

To make the decomposition (10) identifiable one needs an orthogonality constraint:

$$\boldsymbol{E}[\,(\boldsymbol{\beta}^T\vec{\boldsymbol{X}})\,\eta(\vec{\boldsymbol{X}})\,] \;=\; 0.$$

For $\eta(\vec{\boldsymbol{X}})$ as defined in (5), this equality follows from the more general fact that the nonlinearity $\eta(\vec{\boldsymbol{X}})$ is orthogonal to all predictors. Because we will need similar facts for ϵ and δ as well, we state them all at once:

$$\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\eta\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\epsilon\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\delta\,] \;=\; \boldsymbol{0}. \tag{11}$$

Proofs: $\eta \perp \vec{\boldsymbol{X}}$ because it is the population residual of the regression of $\mu(\vec{\boldsymbol{X}})$ on $\vec{\boldsymbol{X}}$ according to (13) below; $\epsilon \perp \vec{\boldsymbol{X}}$ because $\boldsymbol{E}[\vec{\boldsymbol{X}}\epsilon] = \boldsymbol{E}[\vec{\boldsymbol{X}}\boldsymbol{E}[\epsilon|\vec{\boldsymbol{X}}]] = \boldsymbol{0}$. Finally, $\delta \perp \vec{\boldsymbol{X}}$ i because $\delta = \eta + \epsilon$.

As a consequence of the inclusion of an intercept in $\vec{\boldsymbol{X}}$, centering follows as a special case of (11):

$$\boldsymbol{E}[\epsilon] = \boldsymbol{E}[\eta] = \boldsymbol{E}[\delta] = 0. \tag{12}$$

3.4 Error Terms in Econometrics

In econometric models, where predictors are treated as random, there exists a need to specify how error terms relate stochastically to the predictors. While statisticians would assume errors to be independent of predictors, there is a tendency in econometrics to assume orthogonality only: $\boldsymbol{E}[\,\textit{error}\cdot\vec{\boldsymbol{X}}] \;=\; \boldsymbol{0}$. This weaker condition, however, inadvertently permits nonlinearities as part of the "error" because nonlinearities η are indeed uncorrelated with the predictors according to (11), though not independent from them. Thus econometricians' error terms appear to be the total

deviations $\delta = \epsilon + \eta$. The unusual property of δ as "error term" is that generally $\boldsymbol{E}[\,\delta \,|\, \vec{\boldsymbol{X}}] \neq 0$ in the assumption-lean framework, even though $\boldsymbol{E}[\,\delta\,] = 0$ holds always.

The implications may not always be clear, as the following two examples show: Hausman (1978), in an otherwise groundbreaking article, seems to imply that $\boldsymbol{E}[\,error \cdot \vec{\boldsymbol{X}}] = \boldsymbol{0}$ is equivalent to $\boldsymbol{E}[\,error \,|\, \vec{\boldsymbol{X}}] = 0$ (ibid., p.1251, (1.1a)), which it isn't. White's (1980b, p.818) famous article on heteroskedasticity-consistent standard errors also uses the weaker orthogonality assumption, but he is clear that his misspecification test addresses both nonlinearity and heteroskedasticity (ibid., p.823). He is less clear about the fact that "heteroskedasticity-consistent" standard errors are also "nonlinearity-consistent", even though this is spelled out clearly in his lesser-known article on "Using Least Squares to Approximate Unknown Regression Functions" (White 1980a, p.162-3). For this reason the latter article is the more relevant one for us even though it uses an idiosyncratic framework. As nonlinearity is the more consequential model deviation than heteroskedasticity, the synonym "heteroskedasticity-consistent estimator" for the sandwich estimator is somewhat misleading.

What statisticians are more likely to recognize as "error" is ϵ as defined above; yet they will have misgiving as well because ϵ is the deviation not from the fitted approximate model but from the true response surface. We therefore call it "noise" rather than "error." The noise ϵ is *not* independent of the predictors either, but because $\boldsymbol{E}[\epsilon|\vec{\boldsymbol{X}}] = 0$ it enjoys a stronger orthogonality property than the nonlinearity η: $\boldsymbol{E}[\,g(\vec{\boldsymbol{X}})\,\epsilon\,] = 0$ for all $g(\vec{\boldsymbol{X}}) \in L_2(\boldsymbol{P})$. For full independence one would need the property $\boldsymbol{E}[\,g(\vec{\boldsymbol{X}})\,h(\epsilon)\,] = 0$ for all centered $g(\vec{\boldsymbol{X}}), h(\epsilon) \in L_2(\boldsymbol{P})$, which is not generally the case.

Facts:

- *The error ϵ is independent of $\vec{\boldsymbol{X}}$ iff the conditional distribution of ϵ given $\vec{\boldsymbol{X}}$ is the same across predictor space:* $\boldsymbol{E}[f(\epsilon) \,|\, \vec{\boldsymbol{X}}] \overset{\boldsymbol{P}}{=} \boldsymbol{E}[f(\epsilon)] \;\; \forall f(\epsilon) \in L_2$ *(which implies heteroskedasticity).*
- *The nonlinearity η is independent of $\vec{\boldsymbol{X}}$ iff it vanishes:* $\eta \overset{\boldsymbol{P}}{=} 0$.
- *The total deviation δ is independent of $\vec{\boldsymbol{X}}$ iff both the error ϵ is independent of $\vec{\boldsymbol{X}}$ and* $\eta \overset{\boldsymbol{P}}{=} 0$.

As technically trivial as these facts are, they show that *stochastic independence of errors and predictors* is a strong assumption that rules out both nonlinearities and heteroskedasticities. (This form of independence is to be distinguished from the assumption of i.i.d. errors in the linear model where the predictors are fixed.) A hint of unclarity in this regard can be detected even in White (1980b, p.824, footnote 5) when he writes "specification tests ... may detect only a lack of independence between errors and regressors, instead of misspecification." However, "lack of independence" *is* misspecification, which is to first order nonlinearity and to second order heteroskedasticity. What White apparently had in mind are higher order violations of independence. Note that misspecification in the weak sense of violation of orthogonality of errors and predictors is not a meaningful concept: neither nonlinearities nor heteroskedastic noise would qualify as misspecifications as both are orthogonal to the predictors by construction (see Sections 3.2 and 3.3).

4. NON-ANCILLARITY OF THE PREDICTOR DISTRIBUTION

We show in detail that the principle of predictor ancillarity does not hold when models are approximations rather than generative truths. For some background on ancillarity, see Appendix A.

It is clear that the population coefficients $\boldsymbol{\beta}(\boldsymbol{P})$ do not depend on all details of the joint $(Y, \vec{\boldsymbol{X}})$ distribution. For one thing, they are blind to the noise ϵ. This follows from the fact that $\boldsymbol{\beta}(\boldsymbol{P})$ is also

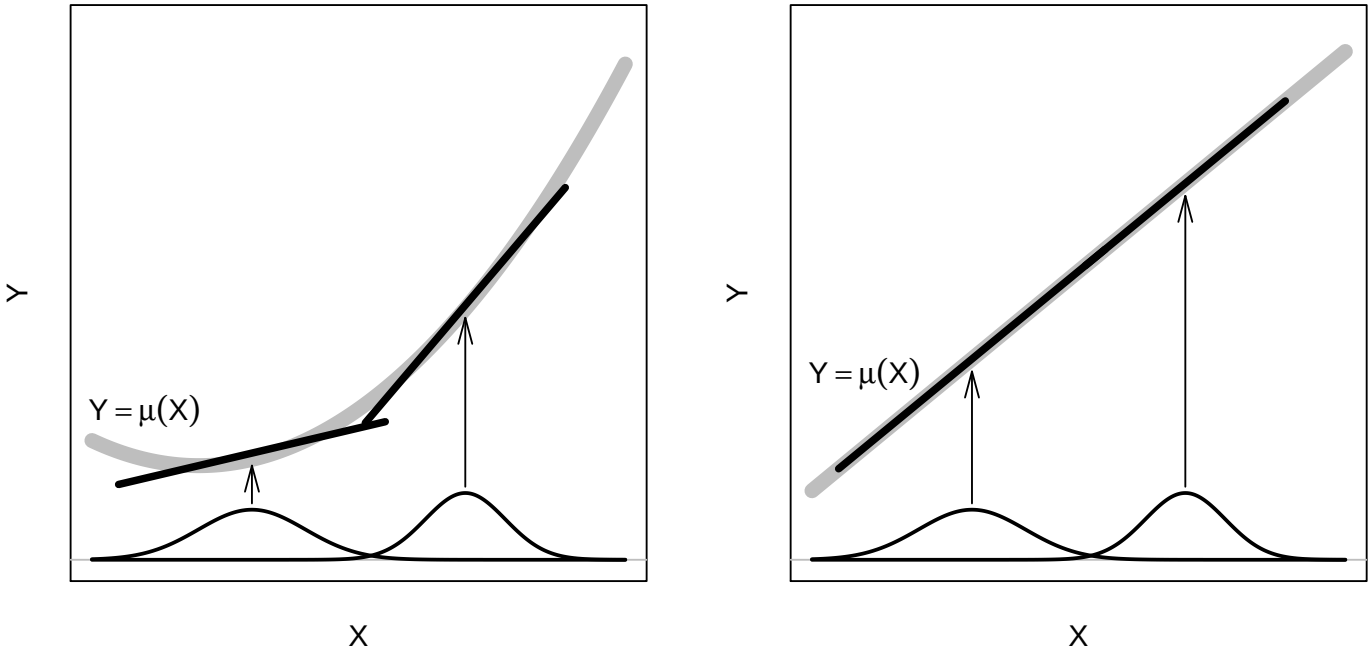

FIG 2. *Illustration of the dependence of the population LS solution on the marginal distribution of the predictors: The left figure shows dependence in the presence of nonlinearity; the right figure shows independence in the presence of linearity.*

the best linear $L_2(\boldsymbol{P})$ approximation to $\mu(\vec{\boldsymbol{X}})$:

$$\boldsymbol{\beta}(\boldsymbol{P}) \;=\; \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \boldsymbol{E}[(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}^T\vec{\boldsymbol{X}})^2] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}\boldsymbol{E}[\,\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})\,]\,. \tag{13}$$

This may be worth spelling out in detail:

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

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

The next and most critical question is whether $\boldsymbol{\beta}(\boldsymbol{P})$ depends on the predictor distribution at all. If the ancillarity argument for predictors is to be believed, the distribution of the predictors should be unrelated to $\boldsymbol{\beta}(\boldsymbol{P})$. The facts, however, are as follows:

Proposition: *The LS functional $\boldsymbol{\beta}(\boldsymbol{P})$ does* **not** *depend on the predictor 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:*

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

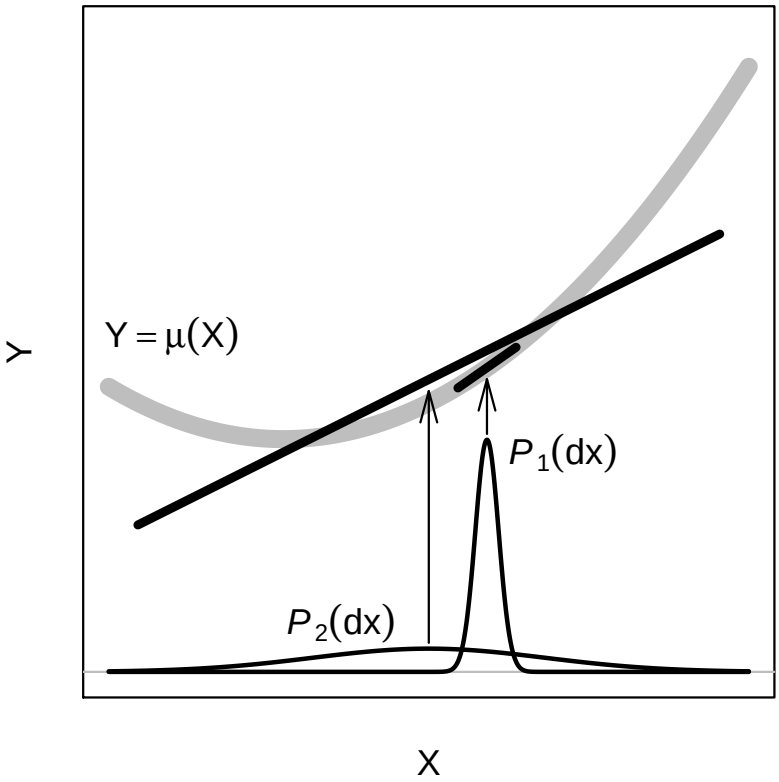

FIG 3. *Illustration of the interplay between predictors' high-density range and nonlinearity: Over the small range of* $\boldsymbol{P}_1$ *the nonlinearity will be undetectable and immaterial for realistic sample sizes, whereas over the extended range of* $\boldsymbol{P}_2$ *the nonlinearity is more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

(For the simple proof details, see Appendix B.2.) 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 predictor 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 proposition is much more easily explained with a graphical illustration: Figure 2 shows single predictor situations with a nonlinear and a linear mean function, respectively, and the same two predictor distributions. The two population LS lines for the two predictor distributions differ in the nonlinear case and they are identical in the linear case. (This observation appears first in White (1980a, p.155-6); to see the correspondence, identify Y with his $g(Z) + \epsilon$.)

The relevance of the proposition is that in the presence of nonlinearity the LS functional $\boldsymbol{\beta}(\boldsymbol{P})$ depends on the predictor distribution, hence the predictors are not ancillary for $\boldsymbol{\beta}(\boldsymbol{P})$. The concept of ancillarity in generative models has things upside down in that it postulates independence of the predictor distribution from the parameters of interest. In a semi-parametric framework where the fitted function is an approximation and the parameters are statistical functionals, the matter presents itself in reverse: It is not the parameters that affect the predictor distribution; rather, it is the predictor distribution that affects the parameters.

The loss of predictor ancillarity has practical implications: Consider two empirical studies that use the same predictor and response variables. If their statistical inferences about $\boldsymbol{\beta}(\boldsymbol{P})$ seem superficially contradictory, there may yet be no contradiction if the response surface is nonlinear and the predictor distributions in the two studies differ: it is then possible that the two studies differ in their targets of estimation, $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$. A difference in predictor distributions in two studies implies in general a difference in the best fitting linear approximation, as illustrated by Figure 2. Differences in predictor distributions can become increasingly complex and harder to detect as the predictor dimension increases.

If at this point one is tempted to recoil from the idea of models as approximations and revert to the idea that models must be well-specified, one should expect little comfort: The very idea of well-specification is a function of the high-density range of predictor distributions because over a small range a model has a better chance of appearing "well-specified" for the simple reason that approximations work better over small ranges . This is illustrated by Figure 3: the narrow range of the predictor distribution $\boldsymbol{P}_1(d\vec{\boldsymbol{x}})$ is the reason why the linear approximation is excellent, that is, the model is very nearly "well specified", whereas the wide range of $\boldsymbol{P}_2(d\vec{\boldsymbol{x}})$ is the reason for the gross "misspecification" of the linear approximation. This is a general issue that cannot be resolved by calls for more "substantive theory" in modeling: Even the best of theories have limited ranges of validity as has been shown by the most successful theories known to science, those of physics.

5. OBSERVATIONAL DATASETS, ESTIMATION, AND CLTS

Turning to estimation from i.i.d. data, it will be shown how the variability in the LS estimate can be asymptotically decomposed into two sources: nonlinearity and noise.

5.1 Notation for Observations Datasets

Moving from populations to samples and estimation, we introduce notation for "observational data", that is, cross-sectional data consisting of i.i.d. cases $(Y_i, X_{i,1}, ..., X_{i,p})$ drawn from a joint multivariate distribution $\boldsymbol{P}(\mathrm{d}y, \mathrm{d}x_1, ..., \mathrm{d}x_p)$ ($i = 1, 2, ..., N$). (Note that White (1980a,b) permits "i.n.i.d." sampling, that is independent but *not* identically distributed observations. His theory imposes technical moment conditions that limit the degree to which the distributions deviate from each other. We use the simpler i.i.d. condition for greater clarity but lesser generality.)

We collect the predictors of case i in a column $(p+1)$-vector $\vec{\boldsymbol{X}}_i = (1, X_{i,1}, ..., X_{i,p})^T$, prepended with 1 for an intercept. We stack the N samples to form random column N-vectors and a random predictor $N \times (p+1)$-matrix:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \cdot\cdot \\ \cdot\cdot \\ Y_N \end{bmatrix}, \qquad \boldsymbol{X}_j = \begin{bmatrix} X_{1,j} \\ \cdot\cdot \\ \cdot\cdot \\ X_{N,j} \end{bmatrix}, \qquad \boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{X}_1, ..., \boldsymbol{X}_p] = \begin{bmatrix} \vec{\boldsymbol{X}}_1^{\,T} \\ \cdots \\ \cdots \\ \vec{\boldsymbol{X}}_N^{\,T} \end{bmatrix}.$$

Similarly we stack the values of the mean function $\mu(\vec{\boldsymbol{X}}_i)$, of the nonlinearity $\eta(\vec{\boldsymbol{X}}_i) = \mu(\vec{\boldsymbol{X}}_i) - \vec{\boldsymbol{X}}_i^{\,T}\boldsymbol{\beta}$, of the noise $\epsilon_i = Y_i - \mu(\vec{\boldsymbol{X}}_i)$, of the total deviations δ_i from linearity, and of the conditional noise

standard deviations $\sigma(\vec{\boldsymbol{X}}_i)$ to form random column N-vectors:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu(\vec{\boldsymbol{X}}1) \\ \cdot\cdot \\ \cdot\cdot \\ \mu(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta(\vec{\boldsymbol{X}}_1) \\ \cdot\cdot \\ \cdot\cdot \\ \eta(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \cdot\cdot \\ \cdot\cdot \\ \epsilon_N \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \cdot\cdot \\ \cdot\cdot \\ \delta_N \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma(\vec{\boldsymbol{X}}_1) \\ \cdot\cdot \\ \cdot\cdot \\ \sigma(\vec{\boldsymbol{X}}_N) \end{bmatrix}. \tag{14}$$

The definitions of $\eta(\vec{\boldsymbol{X}})$, ϵ and δ in (5) translate to vectorized forms:

$$\boldsymbol{\eta} \;=\; \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} \;=\; \boldsymbol{Y} - \boldsymbol{\mu}, \qquad \boldsymbol{\delta} \;=\; \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}. \tag{15}$$

It is important to keep in mind the distinction between population and sample properties. In particular, the N-vectors $\boldsymbol{\delta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are *not* orthogonal to the predictor 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 $\langle \boldsymbol{\delta}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\epsilon}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\eta}, \boldsymbol{X}_j \rangle \neq 0$, even though the associated random variables are orthogonal to X_j in the population: $\boldsymbol{E}[\,\delta X_j] = \boldsymbol{E}[\,\epsilon X_j] = \boldsymbol{E}[\,\eta(\vec{\boldsymbol{X}})X_j] = \;0$.

The **sample linear LS estimate** of $\boldsymbol{\beta}$ is the random column $(p+1)$-vector

$$\hat{\boldsymbol{\beta}} \;=\; (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^T \;=\; \text{argmin}_{\tilde{\boldsymbol{\beta}}} \, \|\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2 \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,. \tag{16}$$

Randomness of $\hat{\boldsymbol{\beta}}$ stems from both the random response $\boldsymbol{Y}$ and the random predictors in $\boldsymbol{X}$. Associated with $\hat{\boldsymbol{\beta}}$ are the following:

the hat or projection matrix:	$\boldsymbol{H} \;=\; \boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T,$
the vector of LS fits:	$\hat{\boldsymbol{Y}} \;=\; \boldsymbol{X}\hat{\boldsymbol{\beta}} \;=\; \boldsymbol{H}\boldsymbol{Y},$
the vector of residuals:	$\boldsymbol{r} \;=\; \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$

The vector $\boldsymbol{r}$ of residuals, which arises from $\hat{\boldsymbol{\beta}}$, is distinct from the vector of total deviations $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, which arises from $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$.

5.2 Decomposition of the LS Estimate According to Noise and Nonlinearity

When the predictors are random the sampling variation of the LS estimate $\hat{\boldsymbol{\beta}}$ can be additively decomposed into two components: one due to noise $\boldsymbol{\epsilon}$ and another due to nonlinearity $\boldsymbol{\eta}$ interacting with randomness of the predictors. This decomposition is a direct reflection of $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$.

In the classical linear models theory, which conditions on $\boldsymbol{X}$, the target of estimation is $\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$. When $\boldsymbol{X}$ is treated as random, the target of estimation is the population LS solution $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$. The term $\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$ is then a random vector that is naturally placed between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; (\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \;+\; (\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta}) \tag{17}$$

This decomposition corresponds to the decomposition $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$ as the following lemma shows.

Definition and Lemma: *We define "Estimation Offsets" or "EOs" for short as follows:*

$$\begin{aligned} &\textit{Total EO}: && \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;&=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\delta}, \\ &\textit{Error EO}: && \hat{\boldsymbol{\beta}} - \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] \;&=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}, \\ &\textit{Nonlinearity EO}: && \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta} \;&=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}. \end{aligned} \tag{18}$$

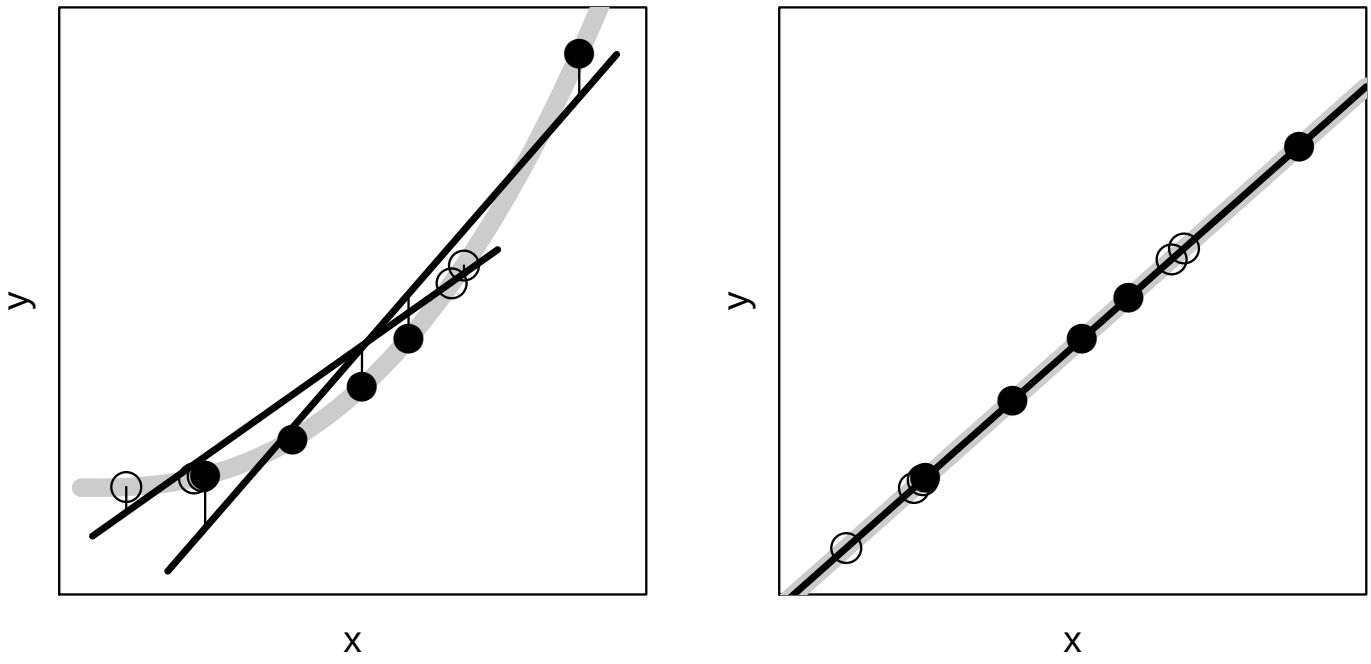

FIG 4. *Noise-less Response: The filled and the open circles represent two "datasets" from the same population. The x-values are random; the y-values are a deterministic function of x: $y = \mu(x)$ (shown in gray).*
Left: The true response $\mu(x)$ is nonlinear; the open and the filled circles have different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

The equations follow from the decompositions (15), $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, and these facts:

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

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

The variance/covariance matrix of $\hat{\boldsymbol{\beta}}$ has a canonical decomposition with regard to conditioning on $\boldsymbol{X}$:

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

This decomposition reflects the estimation decomposition (17) and $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$ in view of (18):

$$\begin{aligned}
\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,] &= \boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\delta}\,], \\
\boldsymbol{E}[\,\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]\,] &= \boldsymbol{E}[\,\boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\,|\boldsymbol{X}]\,], \\
\boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]\,] &= \boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\,].
\end{aligned}$$

In general $\boldsymbol{E}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\,] \neq \boldsymbol{0}$ even though $\boldsymbol{E}[\,\boldsymbol{X}^T\boldsymbol{\eta}\,] = \boldsymbol{0}$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta} \to \boldsymbol{0}$ a.s.

5.3 Random X and Nonlinearity as a Source of Sampling Variation

Linear models theory is largely about sampling variability due to noise $\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]$. The fact that there exists another source of sampling variability is little known: nonlinearity in the presence of

random predictors, as expressed by $\boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]]$ in (5.2). This source can be illustrated in a noise-free situation: Consider a response that is a deterministic but nonlinear function of the predictors: $Y = \eta(\vec{\boldsymbol{X}})$. This is a realistic situation when outputs from expensive deterministic simulation experiments are modeled based on inputs. Assume therefore $\boldsymbol{\epsilon}=\boldsymbol{0}$ but $\boldsymbol{\eta}\neq\boldsymbol{0}$, hence there exists sampling variability in $\hat{\boldsymbol{\beta}}$ which is solely due to the nonlinearity $\boldsymbol{\eta}$: $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}$ in conjunction with the randomness of the predictors — the "conspiracy" in the title of this article.

Figure 4 illustrates the situation with a single-predictor example by showing the LS lines fitted to two "datasets" consisting of $N = 5$ predictor values each. The random differences between datasets cause the fitted line to exhibit sampling variability under nonlinearity (left hand figure), which is absent under linearity (right hand figure). Compare this figure with the earlier Figure 2: mathematically the effects illustrated in both are identical; Figure 2 shows the effect for different populations (theoretical $\boldsymbol{X}$ distributions) while Figure 4 shows it for different datasets (empirical $\boldsymbol{X}$ distributions). Thus nonlinearity creates complications on two interconnected levels: (1) in the definition of the population LS parameter, which becomes dependent on the predictor distribution, and (2) through the creation of sampling variability due to $\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}]$ which becomes a true random vector. (A more striking illustration in the form of an animation 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")`)

The case of a noise-free but nonlinear response is of interest to make another point regarding statistical inference: If classical linear models theory conditions on the predictors and assumes erroneously that the response surface is linear, it is not so that the resulting procedures do "not see" see the sampling variability caused by nonlinearity, but they misinterpret it as due to noise. The consequences of the confusion of noise and nonlinearities for statistical inference will be examined in Section 8.3. This misinterpretation also seeps into the residual bootstrap as it assumes the residuals to originate from exchangeable noise only. By comparison, the pairs bootstrap gets statistical inference right even in the noise-free nonlinear case, at least asymptotically. It receives its justification from the following central limit theorems.

5.4 Assumption-Lean Central Limit Theorems

The three EOs of Section 5.2 arise from the decomposition (5): $\delta = \epsilon + \eta$. The respective CLTs draw on the analogous conditional second moment decomposition $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ (9). The asymptotic variance/covariance matrices have the well-known sandwich form:

Proposition: *The three EOs follow central limit theorems under usual multivariate CLT assumptions:*

$$
\begin{aligned}
N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad &\xrightarrow{\mathcal{D}} \quad \mathcal{N}\Big(\boldsymbol{0},\ \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\Big) \\
N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \quad &\xrightarrow{\mathcal{D}} \quad \mathcal{N}\Big(\boldsymbol{0},\ \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\Big) \qquad (19) \\
N^{1/2}\,(\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta}) \quad &\xrightarrow{\mathcal{D}} \quad \mathcal{N}\Big(\boldsymbol{0},\ \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\eta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\Big)
\end{aligned}
$$

The proof is standard, but an outline for the first case is in Appendix B.3, if only to show how the sandwich form of the asymptotic variance arises.

The center parts of the first two asymptotic sandwich covariances can equivalently be written as

$$\boldsymbol{E}[\, m^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] \;=\; \boldsymbol{E}[\, \delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T], \qquad \boldsymbol{E}[\, \sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] \;=\; \boldsymbol{E}[\, \epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T], \tag{20}$$

which follows from $m^2(\vec{\boldsymbol{X}}) = \boldsymbol{E}[\, \delta^2|\vec{\boldsymbol{X}}]$ and $\sigma^2(\vec{\boldsymbol{X}}) = \boldsymbol{E}[\, \epsilon^2|\vec{\boldsymbol{X}}]$ according to (8) and (9).

The proposition can be specialized in a few ways to cases of partial or complete well-specification:

- **First order well-specification:** When there is no nonlinearity, $\eta(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$, then

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

  The sandwich form of the asymptotic variance/covariance matrix is solely due to heteroskedasticity.

- **First and second order well-specification:** When additionally homoskedasticity holds, $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} \sigma^2$, then

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

  The familiar simplified form is asymptotically valid under first and second order well-specification but without the assumption of Gaussian noise.

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

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

  The sandwich form of the asymptotic variance/covariance matrix is solely due to nonlinearity and random predictors.

6. THE SANDWICH ESTIMATOR AND THE M-OF-N PAIRS BOOTSTRAP

Empirically one observes that standard error estimates obtained from the pairs 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 (30). A closer connection between them will be established below.

6.1 The Plug-In Sandwich Estimator of Asymptotic Variance

According to (19) the asymptotic variance of the LS estimator $\hat{\boldsymbol{\beta}}$ is

$$\boldsymbol{AV}[\hat{\boldsymbol{\beta}}] \;=\; \boldsymbol{E}[\, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\, \boldsymbol{E}[\, \delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\, \boldsymbol{E}[\, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}. \tag{21}$$

The sandwich estimator is then the plug-in version of (21) where δ^2 is replaced by residuals and population expectations $\boldsymbol{E}[...]$ by sample means $\hat{\boldsymbol{E}}[...]$:

$$\begin{aligned}
\hat{\boldsymbol{E}}[\, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] \;&=\; \tfrac{1}{N}\textstyle\sum_{i=1\ldots N} \vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i^T \;&=\; \tfrac{1}{N}\,(\boldsymbol{X}^T\boldsymbol{X}) \\
\hat{\boldsymbol{E}}[\, r^2\, \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] \;&=\; \tfrac{1}{N}\textstyle\sum_{i=1\ldots N} r_i^2\vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i^T \;&=\; \tfrac{1}{N}\,(\boldsymbol{X}^T D_{\boldsymbol{r}}^2\, \boldsymbol{X}),
\end{aligned}$$

where $D_{\boldsymbol{r}}^2$ 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):

$$\begin{aligned} \hat{\boldsymbol{A}}\hat{\boldsymbol{V}}_{sand} &:= \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\hat{\boldsymbol{E}}[\,r^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1} \\ &= N\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,(\boldsymbol{X}^T D_{\boldsymbol{r}}^2\,\boldsymbol{X})\,(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned} \tag{22}$$

The sandwich standard error estimate for the j'th regression coefficient is therefore obtained as

$$\hat{\boldsymbol{SE}}_{sand}[\hat{\beta}_j] := \tfrac{1}{N^{1/2}}(\hat{\boldsymbol{A}}\hat{\boldsymbol{V}}_{sand})_{jj}^{1/2}. \tag{23}$$

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 (23) 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 pairs bootstrap, to which we now turn.

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

To connect the sandwich estimator to its bootstrap counterpart we need the M-of-N bootstrap whereby the *resample size* M is allowed to differ from the sample size N. It is at this point important not to confuse

- M-of-N resampling *with* replacement, and
- M-out-of-N subsampling *without* replacement.

In resampling the resample size M can be any $M < \infty$, whereas for subsampling it is necessary that the subsample size M satisfy $M < N$. We are here concerned with bootstrap resampling, and we will focus on the extreme case $M \gg N$, namely, the limit $M \to \infty$.

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 $\{(\vec{\boldsymbol{X}}_i, Y_i)\}_{i=1\ldots N}$, which is frozen of size N as $M \to \infty$.

Proposition: *For any fixed dataset of size N, there holds a CLT for the M-of-N bootstrap as $M \to \infty$. Denoting by $\boldsymbol{\beta}^*_M$ the LS estimate obtained from a bootstrap resample of size M, we have*

$$M^{1/2}\,(\boldsymbol{\beta}^*_M - \hat{\boldsymbol{\beta}}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\right) \quad (M\to\infty). \tag{24}$$

This is a straight application of the CLT of the previous section to the empirical distribution rather than the actual distribution of the data, where the middle part (the "meat") of the asymptotic formula is based on the empirical counterpart $r_i^2 = (Y_i - \vec{\boldsymbol{X}}_i^T\hat{\boldsymbol{\beta}})^2$ of $\delta^2 = (Y-\vec{\boldsymbol{X}}^T\boldsymbol{\beta})^2$. A comparison of (22) and (24) results in the following:

Corollary: *The sandwich estimator* (22) *is the asymptotic variance estimated by the limit of the M-of-N pairs bootstrap as $M \to \infty$ for a fixed sample of size N.*

As an inferential method the pairs bootstrap is obviously more flexible and richer in possibilities than the sandwich estimator. The latter is limited to providing a standard error estimate assuming approximate normality of the parameter estimate's sampling distribution. The bootstrap distribution, on the other hand, can be used to generate confidence intervals that are often second order correct (the literature on this topic is too rich to list, so we point only to the standard bootstrap reference by Efron and Tibshirani (1994)).

Further connections are mentioned by MacKinnon and White (1985): Some forms of the sandwich estimator were independently derived by Efron (1982, p.18-19) using the infinitesimal jackknife, and by Hinkley (1977) using what he calls a "weighted jackknife". See Weber (1986) for a concise comparison in the fixed-$\boldsymbol{X}$ linear models framework limited to the problem of heteroskedasticity. A richer context for the relation between the jackknife and bootstrap is given by Wu (1986)

7. ADJUSTED PREDICTORS

The adjustment formulas of this section serve to express the slopes of multiple regressions as slopes in simple regressions using adjusted single predictors. The goal is to analyze the discrepancies between asymptotically proper and improper standard errors of regression estimates, and to provide tests that indicate for each predictor whether the linear models standard error is invalidated by "misspecification" (Section 8).

7.1 Adjustment in populations

To express the population LS regression coefficient $\beta_j = \beta_j(\boldsymbol{P})$ as a simple regression coefficient, let the adjusted predictor $X_{j\bullet}$ be defined as the "residual" of the population regression of X_j, used as the response, on all other predictors. In detail, collect all other predictors in the random p-vector $\vec{\boldsymbol{X}}_{-j} = (1, X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)^T$, and let $\boldsymbol{\beta}_{-j\bullet}$ be the coefficient vector from the regression of X_j onto $\vec{\boldsymbol{X}}_{-j}$:

$$\boldsymbol{\beta}_{-j\bullet} \;=\; \mathrm{argmin}_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p}\, \boldsymbol{E}[\,(X_j - \tilde{\boldsymbol{\beta}}^T \vec{\boldsymbol{X}}_{-j})^2] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}_{-j}\vec{\boldsymbol{X}}_{-j}^T]^{-1}\boldsymbol{E}[\,\vec{\boldsymbol{X}}_{-j}X_j]\,.$$

The adjusted predictor $X_{j\bullet}$ is the residual from this regression:

$$X_{j\bullet} \;=\; X_j - \vec{\boldsymbol{X}}_{-j}^T\boldsymbol{\beta}_{-j\bullet}\,. \tag{25}$$

The representation of β_j as a simple regression coefficient is as follows:

$$\beta_j \;=\; \frac{\boldsymbol{E}[\,Y X_{j\bullet}]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]} \;=\; \frac{\boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}})X_{j\bullet}]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]}\,. \tag{26}$$

7.2 Adjustment in samples

To express estimates of regression coefficients as simple regressions, we collect all predictor columns other than $\boldsymbol{X}_j$ in a $N \times p$ random predictor matrix $\boldsymbol{X}_{-j} = (\boldsymbol{1}, ..., \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, ...)$ and define

$$\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} \;=\; \mathrm{argmin}_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p}\, \|\boldsymbol{X}_j - \boldsymbol{X}_{-j}\tilde{\boldsymbol{\beta}}\|^2 \;=\; (\boldsymbol{X}_{-j}^T\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^T\boldsymbol{X}_j\,.$$

Using the hat notation "$\hat{\bullet}$" to denote sample-based adjustment to distinguish it from population-based adjustment "$\bullet$", we write the sample-adjusted predictor as

$$\boldsymbol{X}_{j\hat{\bullet}} \;=\; \boldsymbol{X}_j - \boldsymbol{X}_{-j}\,\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} \;=\; (\boldsymbol{I} - \boldsymbol{H}_{-j})\boldsymbol{X}_j\,. \tag{27}$$

where $\boldsymbol{H}_{\boldsymbol{-j}} = \boldsymbol{X}_{\boldsymbol{-j}}(\boldsymbol{X}_{\boldsymbol{-j}}^T\boldsymbol{X}_{\boldsymbol{-j}})^{-1}\boldsymbol{X}_{\boldsymbol{-j}}^T$ is the associated projection or hat matrix. The j'th slope estimate of the multiple linear regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1, ..., \boldsymbol{X}_p$ can then be expressed in the well-known manner as the slope estimate of the simple linear regression without intercept of $\boldsymbol{Y}$ on $\boldsymbol{X}_{j\hat{\bullet}}$:

$$\hat{\beta}_j \;=\; \frac{\langle \boldsymbol{Y}, \boldsymbol{X}_{j\hat{\bullet}}\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,. \tag{28}$$

In the proofs (see the Appendix) we also need notation for each observation's population-adjusted predictors: $\boldsymbol{X}_{j\bullet} = (X_{1,j\bullet}, ..., X_{N,j\bullet})^T = \boldsymbol{X}_j - \boldsymbol{X}_{\boldsymbol{-j}}\boldsymbol{\beta}_{\boldsymbol{-j\bullet}}$. 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:

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

7.3 Adjustment for Estimation Offsets and Their CLTs

The vectorized formulas for estimation offsets (17) can be written componentwise using adjustment as follows:

$$\begin{aligned}
&\textit{Total EO}: && \hat{\beta}_j - \beta_j &&= \frac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\delta}\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2},\\
&\textit{Error EO}: && \hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] &&= \frac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\epsilon}\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2},\\
&\textit{Nonlinearity EO}: && \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j &&= \frac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\eta}\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}.
\end{aligned} \tag{29}$$

To see these identities directly, note the following, in addition to (28): $\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}] = E[\langle \boldsymbol{\mu}, \boldsymbol{X}_{j\hat{\bullet}}\rangle]/\|\boldsymbol{X}_{j\hat{\bullet}}\|^2$ 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 $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$.

Asymptotic normality can also be expressed for each $\hat{\beta}_j$ separately using population adjustment:

Corollary:

$$\begin{aligned}
N^{1/2}(\hat{\beta}_j - \beta_j) \quad &\stackrel{\mathcal{D}}{\longrightarrow} \quad \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) \;=\; \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,\delta^2 X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right)\\
N^{1/2}(\hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}]) \quad &\stackrel{\mathcal{D}}{\longrightarrow} \quad \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) \;=\; \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,\epsilon^2 X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right)\\
N^{1/2}(\boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j) \quad &\stackrel{\mathcal{D}}{\longrightarrow} \quad \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right)
\end{aligned} \tag{30}$$

These are not new results but reformulations for the components of the vector CLTs (19). The equalities in the first and second case are based on (20). The asymptotic variances of (30) are the subject of next section.

8. ASYMPTOTIC VARIANCES — PROPER AND IMPROPER

The following prepares the ground for an asymptotic comparison of linear models standard errors with correct assumption-lean standard errors. We know the former to be potentially incorrect in the presence of nonlinearity and/or heteroskedasticity, hence a natural question is: by how much can linear models standard errors deviate from valid assumption-lean standard errors? We look for an answer in the asymptotic limit, which frees us from issues related to how the standard errors are estimated.

8.1 Proper Asymptotic Variances in Terms of Adjusted Predictors

The CLTs (30) contain three asymptotic variances, one for the estimate $\hat\beta_j$ and two for the contributions due to noise and due to nonlinearity according to $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta(\vec{\boldsymbol{X}})$. These asymptotic variances are of the same functional form, which suggests using generic notation for all three. We therefore define:

Definition:

$$\boldsymbol{AV}^{(j)}_{lean}[f^2(\vec{\boldsymbol{X}})] \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2} \tag{31}$$

Using $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ from (9), we obtain a decomposition of asymptotic variance suggested by (30):

$$\begin{aligned}
\boldsymbol{AV}^{(j)}_{lean}[m^2(\vec{\boldsymbol{X}})] \quad &= \quad \boldsymbol{AV}^{(j)}_{lean}[\sigma^2(\vec{\boldsymbol{X}}) \quad + \quad \boldsymbol{AV}^{(j)}_{lean}[\eta^2(\vec{\boldsymbol{X}})] \\
\frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2} \quad &= \quad \frac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2} \quad + \quad \frac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}
\end{aligned} \tag{32}$$

8.2 Improper Asymptotic Variances in Terms of Adjusted Predictors

Next we write down an asymptotic form for the conventional standard error estimate from linear models theory in the assumption-lean framework. This asymptotic form will have the appearance of an asymptotic variance but it will generally be improper as its intended domain of validity is the assumption-loaded framework of linear models theory. This "improper" asymptotic variance derives from an estimate $\hat\sigma^2$ of the noise variance, usually $\hat\sigma^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(N–p–1)$. In the assumption-lean framework with both heteroskedastic error variance and nonlinearity, $\hat\sigma^2$ has the following limit for fixed p:

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

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

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \hat\sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \qquad \hat{\boldsymbol{SE}}^2_{lin}[\,\hat\beta_j] \;=\; \frac{\hat\sigma^2}{\|X_{j\hat\bullet}\|^2}. \tag{33}$$

Their scaled limits under lean assumptions are as follows:

$$N\,\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \quad \xrightarrow{\boldsymbol{P}} \quad \boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,]\;\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}^T\,]^{-1}, \qquad N\,\hat{\boldsymbol{SE}}^2_{lin}[\,\hat\beta_j] \quad \xrightarrow{\boldsymbol{P}} \quad \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,]}{\boldsymbol{E}[\,X^2_{j\bullet}\,]}. \tag{34}$$

We call these limits "***improper asymptotic variances***". Again we can use (9) $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ for a decomposition and therefore introduce generic notation where $f^2(\vec{\boldsymbol{X}})$ is a placeholder for any one among $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$:

Definition:

$$\boldsymbol{AV}_{lin}^{(j)}[f^2(\vec{\boldsymbol{X}})] \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,{X_j}_\bullet{}^2]} \tag{35}$$

Hence this the improper asymptotic variance of $\hat{\beta}_j$ and its decomposition:

$$\begin{aligned} \boldsymbol{AV}_{lin}^{(j)}[m^2(\vec{\boldsymbol{X}})] \;&=\; \boldsymbol{AV}_{lin}^{(j)}[\sigma^2(\vec{\boldsymbol{X}})] \;+\; \boldsymbol{AV}_{lin}^{(j)}[\eta^2(\vec{\boldsymbol{X}})] \\ \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,{X_j}_\bullet{}^2]} \;&=\; \frac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,{X_j}_\bullet{}^2]} \;+\; \frac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,{X_j}_\bullet{}^2]} \end{aligned} \tag{36}$$

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

We examine next the discrepancies between proper and improper asymptotic variances by forming their ratio. It will be shown that this ratio can be arbitrarily close to 0 and to ∞. It can be formed separately for each of the versions corresponding to $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$. For this reason we introduce a generic form of the ratio:

Definition: *Ratio of Asymptotic Variances, Proper/Improper.*

$$\boldsymbol{RAV}_j[f^2(\vec{\boldsymbol{X}})] \;:=\; \frac{\boldsymbol{AV}_{lean}^{(j)}[f^2(\vec{\boldsymbol{X}})]}{\boldsymbol{AV}_{lin}^{(j)}[f^2(\vec{\boldsymbol{X}})]} \;=\; \frac{\boldsymbol{E}[f^2(\vec{\boldsymbol{X}}){X_j}_\bullet{}^2]}{\boldsymbol{E}[f^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[{X_j}_\bullet{}^2]} \tag{37}$$

Again, $f^2(\vec{\boldsymbol{X}})$ is a placeholder for each of $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$. The overall $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$ can be decomposed into a weighted average of $\boldsymbol{RAV}_j[\sigma^2(\vec{\boldsymbol{X}})]$ and $\boldsymbol{RAV}_j[\eta^2(\vec{\boldsymbol{X}})]$:

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

$$\begin{aligned} &\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] \;=\; w_\sigma\,\boldsymbol{RAV}_j[\sigma^2(\vec{\boldsymbol{X}})] \;+\; w_\eta\,\boldsymbol{RAV}_j[\eta^2(\vec{\boldsymbol{X}})] \\ &w_\sigma := \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}\,, \qquad w_\eta := \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}\,, \qquad w_\sigma + w_\eta = 1. \end{aligned} \tag{38}$$

Implications of this decomposition will be discussed below. Structurally, the three ratios $\boldsymbol{RAV}_j$ can be interpreted as inner products between the normalized squared random variables

$$\frac{m^2(\vec{\boldsymbol{X}})}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}, \qquad \frac{\sigma^2(\vec{\boldsymbol{X}})}{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]}, \qquad \frac{\eta^2(\vec{\boldsymbol{X}})}{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]}$$

on the one hand, and the normalized squared adjusted predictor

$$\frac{{X_{j\bullet}}^2}{\boldsymbol{E}[{X_{j\bullet}}^2]}$$

on the other hand. These inner products, however, are *not* correlations, and they are *not* bounded by +1; their natural bounds are rather 0 and ∞, both of which can generally be approached to any degree as will be shown in Subsection 8.5.

8.4 The Meaning of *RAV*

The ratio $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$ shows by what multiple the proper asymptotic variance deviates from the improper one:

- If $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] = 1$, then $\hat{\boldsymbol{SE}}_{lin}[\hat{\beta}_j]$ is asymptotically correct;
- if $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] > 1$, then $\hat{\boldsymbol{SE}}_{lin}[\hat{\beta}_j]$ is asymptotically too small/optimistic;
- if $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] < 1$, then $\hat{\boldsymbol{SE}}_{lin}[\hat{\beta}_j]$ is asymptotically too large/pessimistic.

If, for example, $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] = 4$, then, for large sample sizes, the proper standard error of $\hat{\beta}_j$ is about twice as large as the improper standard error of linear models theory. If, however, $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] = 1$, it does not imply that the model is well-specified because heteroskedasticity and nonlinearity can conspire to make $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] = 1$ even though neither $\sigma^2(\boldsymbol{X}) = \text{const}$ nor $\eta(\vec{\boldsymbol{X}}) = 0$; see the decomposition lemma in Subsection 8.3. If, for example, $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}}) = m_0^2$ constant while neither $\sigma^2(\vec{\boldsymbol{X}})$ is constant nor $\eta^2(\vec{\boldsymbol{X}})$ vanishes, then $\boldsymbol{RAV}_j[m_0^2] = 1$ and the linear models standard error is asymptotically correct, yet the model is "misspecified." Well-specification to first and second order, $\eta(\vec{\boldsymbol{X}}) = 0$ and $\sigma^2(\vec{\boldsymbol{X}}) = \sigma_0^2$ constant, is a sufficient but not necessary condition for asymptotic validity of the conventional standard error.

8.5 The Range of *RAV*

As mentioned ***RAV*** ratios can generally vary between 0 and ∞. The following proposition states the technical conditions under which these bounds are sharp. The formulation is generic in terms of $f^2(\vec{\boldsymbol{X}})$ as placeholder for $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$. The proof is in Appendix B.4.

Proposition:
(a) If $X_{j\bullet}$ has unbounded support on at least one side, that is, if $\boldsymbol{P}[{X_{j\bullet}}^2 > t] > 0\ \forall t > 0$, then

$$\sup_f \boldsymbol{RAV}_j[f^2(\vec{\boldsymbol{X}})] = \infty\,. \tag{39}$$

(b) If the closure of the support of the distribution of $X_{j\bullet}$ contains zero (its mean) but there is no pointmass at zero, that is, if $\boldsymbol{P}[{X_{j\bullet}}^2 < t] > 0\ \forall t > 0$ but $\boldsymbol{P}[{X_{j\bullet}}^2 = 0] = 0$, then

$$\inf_f \boldsymbol{RAV}_j[f^2(\vec{\boldsymbol{X}})] = 0\,. \tag{40}$$

As a consequence, it is in general the case that $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$, $\boldsymbol{RAV}_j[\sigma^2(\vec{\boldsymbol{X}})]$ and $\boldsymbol{RAV}_j[\eta^2(\vec{\boldsymbol{X}})]$ can each range between 0 and ∞. (A slight subtlety arises from the constraint imposed on $\eta(\vec{\boldsymbol{X}})$ by orthogonality (11) to the predictors, but it does not invalidate the general fact.)

The proposition involves only some plausible conditions on the distribution of $X_{j\bullet}$, not all of $\vec{\boldsymbol{X}}$. This follows from the fact that the dependence of $\boldsymbol{RAV}_j[f^2(\vec{\boldsymbol{X}})]$ on the distribution of $\vec{\boldsymbol{X}}$ can be reduced to dependence on the distribution of ${X_{j\bullet}}^2$ through conditioning:

$$\boldsymbol{RAV}_j[f^2(\vec{\boldsymbol{X}})] \;=\; \frac{\boldsymbol{E}[f_j^2(X_{j\bullet})\,{X_{j\bullet}}^2]}{\boldsymbol{E}[f_j^2(X_{j\bullet}]\,\boldsymbol{E}[{X_{j\bullet}}^2]} \qquad \text{where} \qquad f_j^2(X_{j\bullet}) \;:=\; \boldsymbol{E}[f^2(\vec{\boldsymbol{X}})\,|\,{X_{j\bullet}}^2]. \tag{41}$$

The problem then boils down to a single-predictor situation in $X = X_{j\bullet}$ which lends itself to graphical illustration. Figure 5 shows a family of functions $f^2(x)$ that interpolates the range of the $\boldsymbol{RAV}$ from 0 to ∞ for $X \sim \mathcal{N}(0,1)$. (Details are in Appendix B.5.)

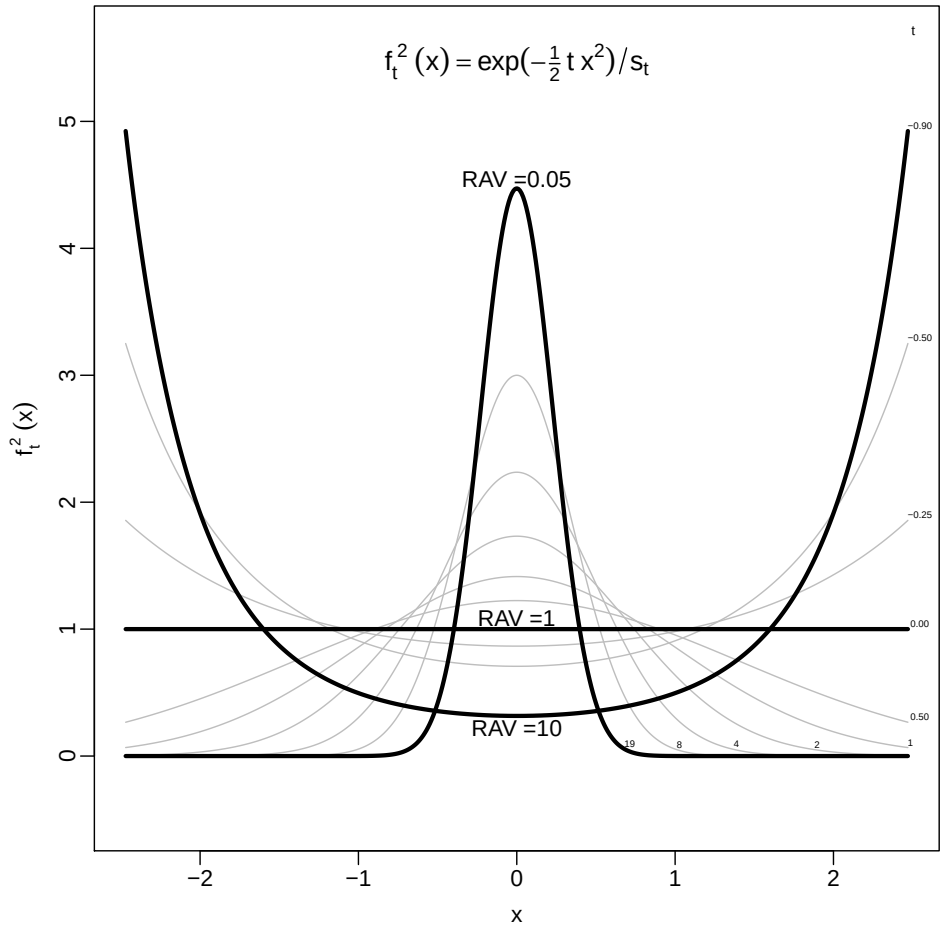

FIG 5. *A family of functions $f_t^2(x)$ that can be interpreted as heteroskedasticities $\sigma^2(X_{j\bullet})$, squared nonlinearities $\eta^2(X_{j\bullet})$, or conditional MSEs $m^2(X_{j\bullet})$: The family interpolates $\boldsymbol{RAV}$ from 0 to ∞ for $x = X_{j\bullet} \sim N(0,1)$. The three solid black curves show $f_t^2(x)$ that result in RAV=0.05, 1, and 10. (See Appendix B.5 for details.)*
$\boldsymbol{RAV} = \infty$ is approached as $f_t^2(x)$ bends ever more strongly in the tails of the x-distribution.
$\boldsymbol{RAV} = 0$ is approached by an ever stronger spike in the center of the x-distribution.

Even though the $\boldsymbol{RAV}$ is not a correlation, it is nevertheless a measure of association between $f_j^2(X_{j\bullet})$ and ${X_{j\bullet}}^2$. Unlike correlations, it exists for $f^2 = \text{const} > 0$ as well, in which case $\boldsymbol{RAV} = 1$. It indicates a positive association between $f^2(\vec{\boldsymbol{X}})$ and ${X_{j\bullet}}^2$ for $\boldsymbol{RAV} > 1$ and a negative association

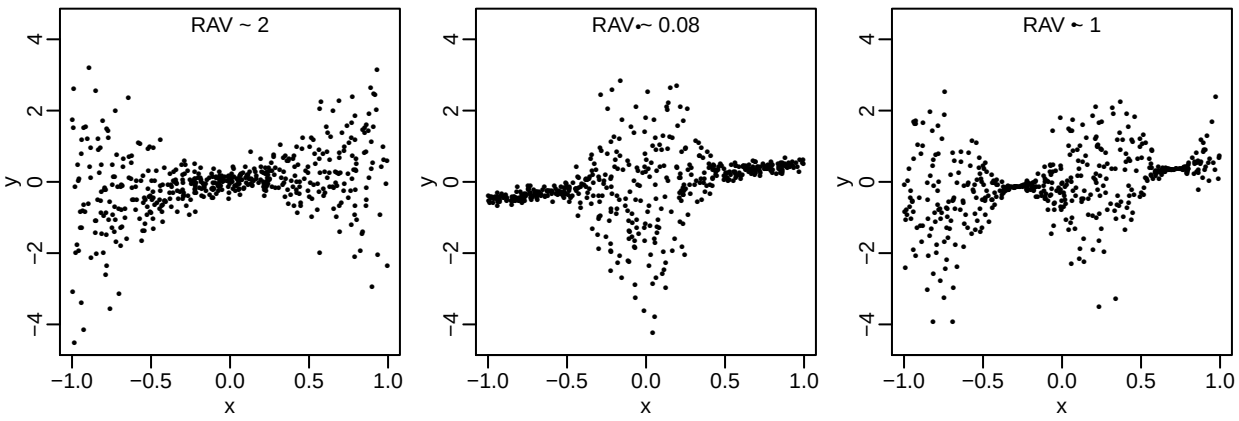

FIG 6. *The effect of heteroskedasticity on the sampling variability of slope estimates: The question is how the misinterpretation of the heteroskedasticities as homoskedastic affects statistical inference.*
*Left: High noise variance in the tails of the predictor distribution elevates the true sampling variability of the slope estimate above the linear models standard error (**RAV**$[\sigma^2(X)] > 1$).*
*Center: High noise variance near the center of the predictor distribution lowers the true sampling variability of the slope estimate below the linear models standard error (**RAV**$[\sigma^2(X)] < 1$).*
*Right: The noise variance oscillates in such a way that the linear models standard error is coincidentally correct (**RAV**$[\sigma^2(X)] = 1$).*

for $\boldsymbol{RAV} < 1$. This is borne out by Figure 5: large values $\boldsymbol{RAV} > 1$ are obtained when $f_j^2(X_{j\bullet})$ is large for $X_{j\bullet}$ far from zero, and small values $\boldsymbol{RAV} < 1$ are obtained when $f^2(X_{j\bullet})$ is large for $X_{j\bullet}$ near zero.

So far we discussed and illustrated the properties of $\boldsymbol{RAV}_j$ in terms of an $\vec{\boldsymbol{X}}$-conditional function $f^2(\vec{\boldsymbol{X}})$ which could be any of $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$. Next we illustrate in terms of potential data situations: Figure 6 shows three heteroskedasticity scenarios and Figure 7 three nonlinearity scenarios. These examples allow us to train our intuitions about the types of heteroskedasticities and nonlinearities that drive the overall $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$. Based on the **RAV** decomposition lemma (38) of Subsection 8.3 according to which $\boldsymbol{RAV}[m^2(\vec{\boldsymbol{X}})]$ is a mixture of $\boldsymbol{RAV}[\sigma^2(\vec{\boldsymbol{X}})]$ and $\boldsymbol{RAV}[\eta^2(\vec{\boldsymbol{X}})]$, we can state the following:

- Heteroskedasticities $\sigma^2(\vec{\boldsymbol{X}})$ with large average variance $\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}}) \,|\, {X_{j\bullet}}^2]$ in the tail of ${X_{j\bullet}}^2$ imply an upward contribution to the overall $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$; heteroskedasticities with large average variance concentrated near ${X_{j\bullet}}^2 = 0$ imply a downward contribution to the overall $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$.
- Nonlinearities $\eta^2(\vec{\boldsymbol{X}})$ with large average values $\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}}) \,|\, {X_{j\bullet}}^2]$ in the tail of ${X_{j\bullet}}^2$ imply an upward contribution to the overall $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$; nonlinearities with large average values concentrated near ${X_{j\bullet}}^2 = 0$ imply a downward contribution to the overall $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$.

These facts also suggest the following: in practice, large values $\boldsymbol{RAV}_j > 1$ are generally more likely than small values $\boldsymbol{RAV}_j < 1$ because both large conditional variances and nonlinearities are often more pronounced in the extremes of predictor 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 lemma (38) that for fixed relative contributions $w_\sigma > 0$ and $w_\eta > 0$ either of $\boldsymbol{RAV}_j[\sigma^2(\vec{\boldsymbol{X}})]$ or $\boldsymbol{RAV}_j[\eta^2(\vec{\boldsymbol{X}})]$ is able to single-handedly pull $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$ to $+\infty$, whereas

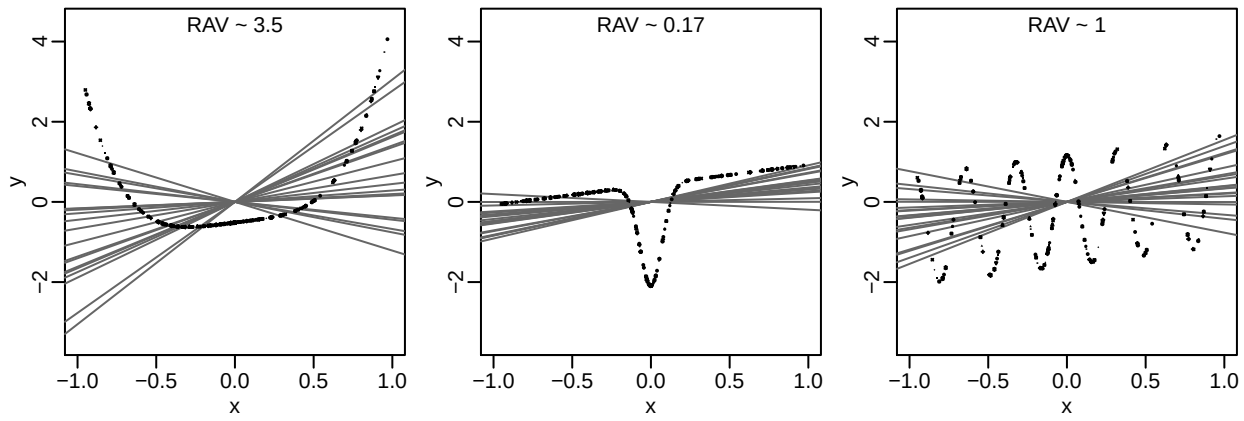

FIG 7. *The effect of nonlinearities on the sampling variability of slope estimates: The three plots show three different noise-free nonlinearities; each plot shows for one nonlinearity 20 overplotted datasets of size $N = 10$ and their fitted lines through the origin. The question is how the misinterpretation of the nonlinearities as homoskedastic random errors affects statistical inference.*
*Left: Strong nonlinearity in the tails of the predictor distribution elevates the true sampling variability of the slope estimate above the linear models standard error (**RAV**$[\eta^2(X)] > 1$).*
*Center: Strong nonlinearity near the center of the predictor distribution lowers the true sampling variability of the slope estimate below the linear models standard error (**RAV**$[\eta^2(X)] < 1$).*
*Right: An oscillating nonlinearity mimics homoskedastic random error to make the linear models standard error coincidentally correct (**RAV**$[\eta^2(X)] = 1$).*

both have to be close to zero to pull $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$ toward zero. These considerations are of course no more than heuristics and practical common sense, but they may be the best we can hope for to understand the prevalence of situations in which the linear models standard error is too small.

9. THE SANDWICH ESTIMATOR IN ADJUSTED FORM AND A *RAV* TEST

The goal is to write the sandwich estimator of standard error in adjustment form and use it to estimate the ***RAV*** with plug-in for use as a test to decide whether the standard error of linear models theory is adequate. In adjustment form we obtain one test per predictor variable. These tests belong in the class of "misspecification tests" for which there exists a literature in econometrics starting with Hausman (1978) and continuing with White (1980a,b; 1981; 1982) and others. The tests of Hausman and White are largely global rather than coefficient-specific, which ours is. Test proposed here has similarities to White's (1982, Section 4) "information matrix test" as it compares two types of information matrices globally, while we compare two types of standard errors one coefficient at a time. The parameter-specific tests of White (1982, Section 5), however, take a different approach altogether: they compare two types of coefficient estimates rather than standard error estimates. The test procedures proposed here have a simplicity and flexibility that may be missing in the extant literature. The flexibility arises from being able to exclude normality of noise from the null hypothesis, which we find important as otherwise most misspecification tests respond to non-normality much of the time rather than nonlinearity and heteroskedasticity.

9.1 The Adjustment Form of the Sandwich Estimator and the $R\hat{A}V_j$ Statistic

To begin with, the adjustment versions of the asymptotic variances in the CLTs (30) can be used to rewrite the sandwich estimator by replacing expectations $\boldsymbol{E}[...]$ with means $\hat{\boldsymbol{E}}[...]$, the population parameter $\boldsymbol{\beta}$ with its estimate $\hat{\boldsymbol{\beta}}$, and population adjustment $X_{j\bullet}$ with sample adjustment $X_{j\hat{\bullet}}$:

$$\hat{\boldsymbol{AV}}^{(j)}_{sand} \;=\; \frac{\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2]}{\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2]^2} \;=\; N\,\frac{\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^4} \tag{42}$$

The squaring of N-vectors is meant to be coordinate-wise. Formula (42) is not a new estimator of asymptotic variance; rather, it is an algebraically equivalent re-expression of the diagonal elements of $\hat{\boldsymbol{AV}}_{sand}$ in (22) above: $\hat{\boldsymbol{AV}}^{(j)}_{sand} = (\hat{\boldsymbol{AV}}_{sand})_{j,j}$. The sandwich standard error estimate (23) can therefore be written as follows:

$$\hat{\boldsymbol{SE}}_{sand}(\hat{\beta}_j) \;=\; \frac{\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle^{1/2}}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}. \tag{43}$$

The usual standard error estimate from linear models theory is (33):

$$\hat{\boldsymbol{SE}}_{lin}(\hat{\beta}_j) \;=\; \frac{\hat{\sigma}}{\|\boldsymbol{X}_{j\hat{\bullet}}\|} \;=\; \frac{\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|}{(N\text{–}p\text{–}1)^{1/2}\,\|\boldsymbol{X}_{j\hat{\bullet}}\|}\,. \tag{44}$$

In order to translate $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$ into a practically useful diagnostic, an obvious first attempt would be forming the ratio $\hat{\boldsymbol{SE}}_{sand}(\hat{\beta}_j)/\hat{\boldsymbol{SE}}_{lin}(\hat{\beta}_j)$, squared. However, $\hat{\boldsymbol{SE}}_{lin}(\hat{\beta}_j)$ has been corrected for fitted degrees of freedom, whereas $\hat{\boldsymbol{SE}}_{sand}(\hat{\beta}_j)$ has not. For greater comparability one would either correct the sandwich estimator with a factor $(N/(N\text{–}p\text{–}1))^{1/2}$ (MacKinnon and White 1985) or else "uncorrect" $\hat{\boldsymbol{SE}}_{lin}(\hat{\beta}_j)$ by replacing $N\text{–}p\text{–}1$ with N in the variance estimate $\hat{\sigma}^2$. Either way one obtains the natural plug-in estimate of $\boldsymbol{RAV}_j$:

$$\boldsymbol{R\hat{A}V}_j \;:=\; N\,\frac{\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\,\|\boldsymbol{X}_{j\hat{\bullet}}\|^2} \;=\; \frac{\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2\,]}{\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2\,]\,\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2\,]}\,. \tag{45}$$

This diagnostic quantity can be used as a test statistic, as will be shown next. The functional form of $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$ and its estimate $\boldsymbol{R\hat{A}V}_j$ illuminates a remark by White (1982) on his "Information Matrix Test for Misspecification" for general ML estimation: "In the linear regression framework, the 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" (ibid., p.12). We know now what function of the predictors actually matters for judging the effects of misspecification on inference for a particular regression coefficient: it is the squared adjusted predictor and its association with the squared total deviations as estimated by residuals.

9.2 A $R\hat{A}V_j$ Test

There exist several ways to generate inference based on the $\boldsymbol{R\hat{A}V}_j$, two of which we discuss in this section, but only one of which can be recommended in practice. We start with an asymptotic result that would be expected to yield approximately valid retention intervals under a null hypothesis of well-specification.

Proposition: *If the total deviations δ_i are independent of $\vec{\boldsymbol{X}}_i$ (not assuming normality of δ_i) we have:*

$$N^{1/2}\,(\boldsymbol{R\hat{A}V}_j - 1) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0,\, \frac{\boldsymbol{E}[\,\delta^4]}{\boldsymbol{E}[\,\delta^2]^2}\,\frac{\boldsymbol{E}[{X_{j\bullet}}^4]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}\;-\;1)\right) \tag{46}$$

If one assumes $\delta_i \sim \mathcal{N}(0,\sigma^2)$, then the asymptotic variance simplifies using $\boldsymbol{E}[\,\delta^4]/\boldsymbol{E}[\,\delta^2]^2 = 3$.

As always we ignore moment conditions among the assumptions. A proof outline is in Appendix B.6.

According to (46) it is the kurtoses (= the standardized fourth moments - 3) of total deviation δ and of the adjusted predictor $X_{j\bullet}$ that drive the asymptotic variance of $\boldsymbol{R\hat{A}V}_j$ under the null hypothesis. We note the following facts:

1. Because standardized fourth moments are always ≥ 1 by Jensen's inequality, the asymptotic variance is ≥ 0, as it should be. The minimal standardized fourth moment of $+1$ is attained by a two-point distribution symmetric about 0. Thus a zero asymptotic variance of $\boldsymbol{R\hat{A}V}_j$ is achieved when both the total deviations δ_i and the adjusted predictor $X_{i,j\bullet}$ have two-point distributions.
2. The larger the kurtosis of δ or $X_{j\bullet}$, the less likely it is that first and second order model misspecification can be detected because the larger the asymptotic standard errors will be. It is an important fact that elevated kurtosis of δ and $X_{j\bullet}$ obscures nonlinearity and heteroskedasticity. Yet, if such misspecification can be detected in spite of elevated kurtoses, it is news worth knowing.
3. A test of the stronger hypothesis that includes normality of δ is obtained by setting $\boldsymbol{E}[\delta^4]/\boldsymbol{E}[\delta^2]^2 = 3$ rather than estimating it. However, the resulting test turns into a non-normality test much of the time. As non-normality can be diagnosed separately with normality tests or normal quantile plots of the residuals, we recommend keeping normality out of the null hypothesis and test independence of δ and $X_{j\bullet}$ alone.

The asymptotic result of the proposition provides insights, but unfortunately it is in our experience not suitable for practical application. The standard procedure would be to estimate the asymptotic null variance of $\boldsymbol{R\hat{A}V}_j$, rescale to sample size N, and use it to form a retention interval around the null value $\boldsymbol{RAV}_j = 1$. The problem is that the null distribution of $\boldsymbol{R\hat{A}V}_j$ in finite datasets can be non-normal in such a way that is not easily overcome by obvious tools such as logarithmic transformations.

Not all is lost, however, because non-asymptotic simulation-based approaches to inference exist for the type of null hypothesis in question. Because the null hypothesis is independence between the total deviation δ and the adjusted predictor $X_{j\bullet}$, a permutation test offers itself. To this end it is necessary that $N \gg p$, and the test will not be exact. The reason is that one needs to estimate the total deviations δ_i with residuals r_i and the population adjusted predictor values $X_{i,j\bullet}$ with sample adjusted predictor values $X_{i,j\hat{\bullet}}$. This test is for the weak hypothesis that does not include normality of δ_i and therefore permits general (centered) noise distributions. A retention interval should be formed directly from the $\alpha/2$ and $1-\alpha/2$ quantiles of the permutation distribution. Quantile-based intervals can be asymmetric according to the skewness and other idiosyncrasies of the permutation distribution. Computations inside the permutation simulation are cheap: Once standardized squared vectors $\boldsymbol{r}^2/\|\boldsymbol{r}\|^2$ and $\boldsymbol{X}_{j\hat{\bullet}}/\|\boldsymbol{X}_{j\hat{\bullet}}\|^2$ are formed, a draw from the conditional null distribution of $\boldsymbol{R\hat{A}V}_j$

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

TABLE 3

Permutation Inference for $\boldsymbol{\hat{RAV}}_j$ in the LA Homeless Data (10,000 permutations).

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{sand}$	$\boldsymbol{\hat{RAV}}_j$	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 4

Permutation Inference for $\boldsymbol{\hat{RAV}}_j$ in the Boston Housing Data (10,000 permutations).

is obtained by randomly permuting one of the vectors and forming the inner product with the other vector. Finally, the approximate permutation distributions can be readily used to diagnose the non-normality of the conditional null using normal quantile plots (see Appendix C for examples).

Tables 3 and 4 show the results for the two datasets of Section 2. Values of $\boldsymbol{\hat{RAV}}_j$ that fall outside the middle 95% range of their permutation null distributions are marked with asterisks. Surprisingly, in the LA Homeless data of Table 3 the values of approximately 2 for the $\boldsymbol{\hat{RAV}}_j$ of "PercVacant" and "PercCommercial" are not statistically significant.

10. THE MEANING OF REGRESSION SLOPES IN THE PRESENCE OF NONLINEARITY

An objection against using linear fits in the presence of nonlinearities is that slopes lose their common interpretation: no longer is β_j the average difference in Y associated with a unit difference in X_j at fixed levels of all other X_k. Yet, there exists a simple alternative interpretation that is valid and intuitive even in the presence of nonlinearities, both for the parameters of the population and their estimates from samples: slopes are weighted averages of case-wise slopes or pairwise slopes. This holds for simple linear regression and also for multiple linear regression for each predictor after linearly adjusting it for all other predictors. This is made precise as follows:

- **Sample estimates**: In a multiple regression based on a sample of size N, consider the LS estimate $\hat{\beta}_j$: this is the empirical simple regression slope through the origin with regard to the empirically adjusted predictor $X_{j\bullet}$ (for $j \neq 0$ as we only consider actual slopes, not the intercept, but assume the presence of an intercept). To simplify notation we write $(x_1, ..., x_N)^T$

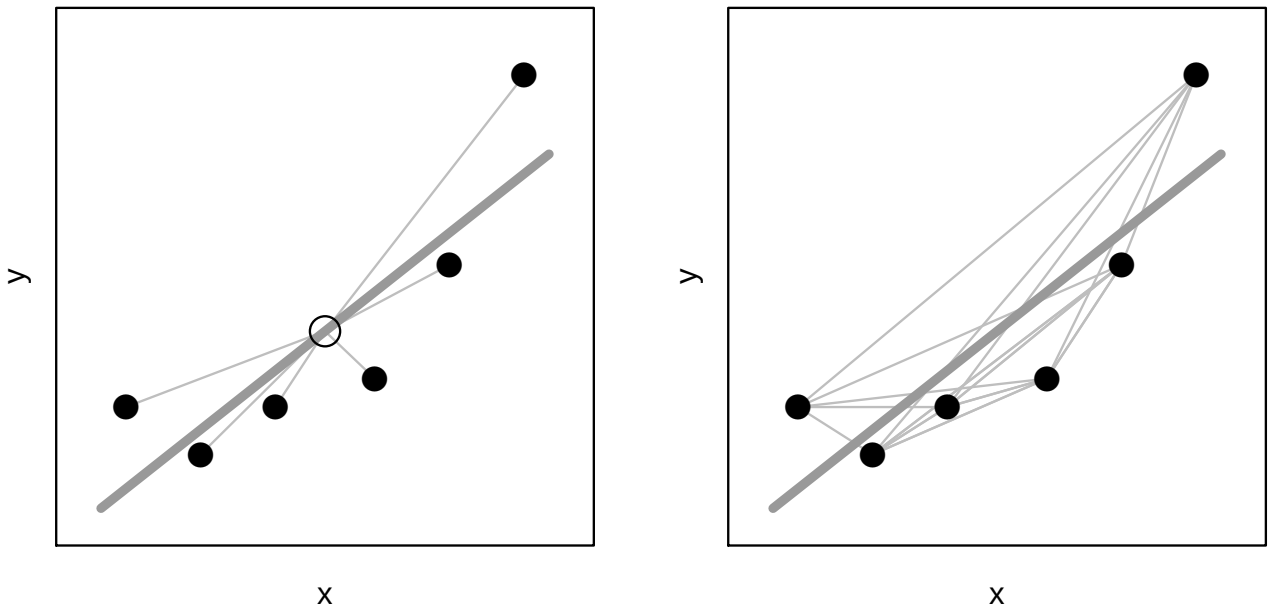

FIG 8. *Case-wise and pairwise average weighted slopes illustrated: Both plots show the same six points (the "cases") as well as the LS line fitted to them (fat gray). The left hand plot shows the case-wise slopes from the mean point (open circle) to the six cases, while the right hand plot shows the pairwise slopes between all 15 pairs of cases. The LS slope is a weighted average of the case-wise slopes on the left according to* (47)*, and of the pairwise slopes on the right according to* (48)*.*

for $X_{j\bullet}$, as well as $(y_1, ..., y_N)^T$ for the response vector $\boldsymbol{Y}$ and $\hat{\beta}$ for the LS estimate $\hat{\beta}_j$. Then the representation of $\hat{\beta}$ as a weighted average of case-wise slopes is

$$\hat{\beta} \;=\; \sum_i w_i\, b_i\,, \quad \text{where} \quad b_i \;:=\; \frac{y_i}{x_i} \quad \text{and} \quad w_i \;:=\; \frac{x_i^2}{\sum_{i'} x_{i'}^2} \tag{47}$$

are case-wise slopes and weights, respectively.
The representation of $\hat{\beta}$ as a weighted average of pairwise slopes is

$$\hat{\beta} \;=\; \sum_{ik} w_{ik}\, b_{ik}\,, \quad \text{where} \quad b_{ik} \;:=\; \frac{y_i - y_k}{x_i - x_k} \quad \text{and} \quad w_{ik} \;:=\; \frac{(x_i - x_k)^2}{\sum_{i'k'} (x_{i'} - x_{k'})^2} \tag{48}$$

are pairwise slopes and weights, respectively. The summations can be over $i \neq k$ or $i < k$. See Figure 8 for an illustration.

- **Population parameters**: In a population multiple regression, consider the slope parameter β_j of the predictor variable X_j. It is also the simple regression slope through the origin with regard to the population-adjusted predictor $X_{j\bullet}$, where again we consider only actual slopes, $j \neq 0$, but assume the presence of an intercept. We now write X instead of $X_{j\bullet}$ and β instead of β_j. The population regression is thus reduced to a simple regression through the origin.
  The representation of β as a weighted average of case-wise slopes is

$$\beta = \boldsymbol{E}[\, W\, B\,], \quad \text{where} \quad B := \frac{Y}{X} \quad \text{and} \quad W := \frac{X^2}{\boldsymbol{E}[\, X^2\,]}$$

are case-wise slopes and case-wise weights, respectively.

For the representation of β as a weighted average of pairwise slopes we need two independent copies (X, Y) and (X', Y') of the predictor and response:

$$\beta \;=\; \boldsymbol{E}[\, W\, B\,] \quad \text{where} \quad B \;:=\; \frac{Y - Y'}{X - X'} \quad \text{and} \quad W \;:=\; \frac{(X - X')^2}{\boldsymbol{E}[\,(X - X')^2\,]}$$

are pairwise slopes and weights, respectively.

These formulas provide intuitive interpretations of regression slopes that are valid without the first order assumption of linearity of the response as a function of the predictors. They support the intuition that, even in the presence of a nonlinearity, a linear fit can be used to infer the overall direction of the association between the response and the predictors.

The above formulas were used and modified to produce alternative slope estimates by Gelman and Park (2008), also 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 from pairs to tuples of size $r \geq p+1$ 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.

11. SUMMARY

In this article we compared statistical inference from classical linear models theory with inference from econometric theory. The major differences are that the former is a finite-sample theory that relies on strong assumptions and treats the predictors as fixed even when they are random, whereas the latter uses asymptotic theory that relies on few assumptions and treats the predictors as random. On a practical level, inferences differ in the type of standard error estimates they use: linear models theory is based on the "usual" standard error which is a scaled version of the noise standard deviation, whereas econometric theory is based on the so-called "sandwich estimator" of standard error which derives from an assumption-lean asymptotic variance. In comparing and contrasting the two modes of statistical inference we observe the following:

- As econometric theory does not assume the correctness of the linearity and homoskedasticity assumptions of linear models theory, a new interpretation of the targets of estimation is needed: Linear fits estimate the best linear approximation to a usually nonlinear response surface.
- If statisticians are willing to buy into this semi-parametric view of linear regression, they will accept sandwich-based inference as asymptotically correct. — If they are unwilling to go down this route, they must have strong belief in the correctness of their models and/or rely on diagnostic methodology to ascertain that linearity and homoskedasticity assumptions are not violated in ways that affect "usual" statistical inference.
- While regression is rich in model diagnostics, a more targeted approach in this case may be based on misspecification tests which are well-established in econometrics. We described one such test which permits testing the adequacy of the linear models standard error, one coefficient at a time.
- The discrepancies between standard errors from assumption-rich linear models theory and assumption-lean econometric theory can be of arbitrary magnitude in the asymptotic limit,

but real data examples indicate discrepancies by a factors of 2 to be common. This is obviously relevant because such factors can change a t-statistic from significant to insignificant and vice versa.

- The pairs bootstrap is seen to be an alternative to the sandwich estimate of standard error. The latter is the asymptotic limit in the M-of-N bootstrap as $M \to \infty$.

Assumption lean inference is not without its problems. A major issue is its non-robustness: compared to the standard error from linear models theory the sandwich standard error relies on higher order moments. The non-robustness is fundamentally a consequence of the LS method, which may suggest that solutions should be obtained through a revival of robustness theory.

REFERENCES

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

[2] Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*, Princeton: Princeton University Press.

[3] Belsley, D. A.., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity.* Wiley series in probability and mathematical statistics. Hoboken, NJ: John Wiley & Sons, Inc.

[4] Berk, R. A., 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.)

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

[6] 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.

[7] Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear Smoothers and Additive Models (with discussions and rejoinder). *The Annals of Statistics*, **17** (2), 453–555.

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

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

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

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

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

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

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

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

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

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

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

[19] 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.

[20] 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.

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

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

[23] 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.

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

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

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

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

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

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

APPENDIX A: ANCILLARITY

The facts as layed out in Section 4 amount to an argument against conditioning on predictors in regression. The justification for conditioning derives from an ancillarity argument according to which the predictors, 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 predictors form an ancillary. To see this we need an extended definition of ancillarity that includes nuisance parameters. The ingredients and conditions are as follows:

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

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

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

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

When applied to regression, the principle implies that in *any* regression model the predictors, 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 predictors. (The predictor 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 B: PROOFS

B.1 Proof of the Lemma in Section 3.4

- Noise ϵ: Assuming constancy of the conditional distribution we obtain independence of the noise as follows:

$$\boldsymbol{E}[f(\epsilon)g(\vec{\boldsymbol{X}})] = \boldsymbol{E}[\boldsymbol{E}[f(\epsilon)|\vec{\boldsymbol{X}}]g(\vec{\boldsymbol{X}})] = \boldsymbol{E}[\boldsymbol{E}[f(\epsilon)]g(\vec{\boldsymbol{X}})] = \boldsymbol{E}[f(\epsilon)]\boldsymbol{E}[g(\vec{\boldsymbol{X}})]$$

  Conversely, if the conditional distribution of the noise is not constant, there exists $f(\epsilon)$ such that $\boldsymbol{E}[f(\epsilon)|\vec{\boldsymbol{X}}] > \boldsymbol{E}[f(\epsilon)]$ for $\vec{\boldsymbol{X}} \in A$ for some A with $\boldsymbol{P}[A] > 0$. Let $g(\vec{\boldsymbol{X}}) = 1_A(\vec{\boldsymbol{X}})$, and it follows $\boldsymbol{E}[f(\epsilon)g(\vec{\boldsymbol{X}})] > \boldsymbol{E}[f(\epsilon)]\,\boldsymbol{E}[g(\vec{\boldsymbol{X}})]$.

- Nonlinearity η: The conditional distribution of η given $\vec{\boldsymbol{X}}$ is a point mass. The same argument as for noise applies, but restricted to point masses. Because $\boldsymbol{E}[\eta] = 0$ (due to the presence of an intercept) the point masses must be at zero.
- Total deviation $\delta = \epsilon + \eta$: Again, the conditional distribution must be identical across predictor space, which results in both of the previous cases.

B.2 Proofs of the Proposition in Section 4

The linear case is trivial: if $\mu_0(\vec{\boldsymbol{X}})$ is linear, that is, $\mu_0(\vec{\boldsymbol{x}}) = \boldsymbol{\beta}^T \vec{\boldsymbol{x}}$ for some $\boldsymbol{\beta}$, then $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}$ irrespective of $\boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}})$ according to (13). The nonlinear case is proved as follows: For any set of points $\vec{\boldsymbol{x}}_1, \ldots \vec{\boldsymbol{x}}_{p+1} \in \mathbb{R}^{p+1}$ in general position and with 1 in the first coordinate, there exists a unique linear function $\boldsymbol{\beta}^T \vec{\boldsymbol{x}}$ 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 $\boldsymbol{\beta}_1^T \vec{\boldsymbol{x}}$ and $\boldsymbol{\beta}_2^T \vec{\boldsymbol{x}}$ 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 $\boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 = \boldsymbol{\beta}(\boldsymbol{P}_2)$.

B.3 Proof Outline of Asymptotic Normality, Proposition of Section 5.4

Using $\boldsymbol{E}[\,\delta \vec{\boldsymbol{X}}\,] = \boldsymbol{0}$ from (12) we have:

$$
\begin{aligned}
N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad &= \quad \left(\tfrac{1}{N}\,\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \left(\tfrac{1}{N^{1/2}}\,\boldsymbol{X}^T \boldsymbol{\delta}\right) \\
&= \quad \left(\tfrac{1}{N}\,\sum \vec{\boldsymbol{X}}_i \vec{\boldsymbol{X}}_i^T\right)^{-1} \left(\tfrac{1}{N^{1/2}}\,\sum \vec{\boldsymbol{X}}_i\,\delta_i\right) \\
&\xrightarrow{\mathcal{D}} \quad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{E}[\,\delta^2 \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\right) \\
&= \quad \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\delta^2 \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\right)\,,
\end{aligned}
$$

B.4 Proof of the Proposition of Section 8.5

An important difference between $\eta^2(\vec{\boldsymbol{X}})$ and $\sigma^2(\vec{\boldsymbol{X}})$ is that nonlinearities are constrained by orthogonalities to the predictors, whereas conditional noise variances are not.

Consider first nonlinearities $\eta(\vec{\boldsymbol{X}})$: We construct a one-parameter family of nonlinearities $\eta_t(\vec{\boldsymbol{X}})$ for which $\sup_t \boldsymbol{RAV}_j[\eta_t^2] = \infty$ and $\inf_t \boldsymbol{RAV}_j[\eta_t^2] = 0$. Generally in the construction of examples, it must be kept in mind that nonlinearities are orthogonal to (adjusted for) all other predictors: $\boldsymbol{E}[\eta(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}] = \boldsymbol{0}$. To avoid uninsightful complications arising from adjustment due to complex dependencies among the predictors, we construct an example for simple linear regression with a single predictor $X_1 = X$ and an intercept $X_0 = 1$. W.l.o.g. we will further assume that X_1 is centered (population adjusted for X_0, so that $X_{1\bullet} = X_1$) and standardized. In what follows we write X instead of X_1, and the assumptions are $\boldsymbol{E}[X] = 0$ and $\boldsymbol{E}[X^2] = 1$.

Proposition: *Define a one-parameter family of nonlinearities as follows:*

$$
\eta_t(X) \;=\; \frac{1_{[|X|>t]} - p(t)}{\sqrt{p(t)(1-p(t))}}\,, \qquad \textit{where} \quad p(t) := \boldsymbol{P}[|X| > t]\,. \tag{49}
$$

We assume that $p(t) > 0\ \forall t > 0$. (We have $1 - p(t) > 0$ for sufficiently large t.) Assume further that the distribution of X is symmetric about 0, so that $\boldsymbol{E}[\eta_t(X)\,X] = 0$. Then we have:

$\lim_{t\uparrow\infty} \boldsymbol{RAV}[\eta_t^2] = \infty$;
$\lim_{t\downarrow 0} \boldsymbol{RAV}[\eta_t^2] = 0$ *if the distribution of* X *has no atom at the origin:* $\boldsymbol{P}[X=0]=0$.

By construction these nonlinearities are centered and standardized, $\boldsymbol{E}[\eta_t(X)] = 0$ and $\boldsymbol{E}[\eta_t(X)^2] = 1$. They are also orthogonal to X, $\boldsymbol{E}[\eta_t(X)X] = 0$, due to the assumed symmetry of the distribution of X, $P[X > t] = P[X < -t]$, and the symmetry of the nonlinearities, $\eta_t(-X) = \eta_t(X)$.

Consider next heteroskedastic noise variances $\sigma^2(\vec{\boldsymbol{X}})$: The above construction for nonlinearities can be re-used. As with nonlinearities, for $\boldsymbol{RAV}[\sigma_t^2(X)]$ to rise with no bound, the conditional noise variance $\sigma_t^2(X)$ needs to place its large values in the unbounded tail of the distribution of X. For $\boldsymbol{RAV}[\sigma_t^2(X)]$ to reach down to zero, $\sigma_t^2(X)$ needs to place its large values in the center of the distribution of X.

Proposition: *Define a one-parameter family of heteroskedastic noise variances as follows:*

$$\sigma_t^2(X) \;=\; \frac{(1_{[|X|>t]} - p(t))^2}{p(t)(1-p(t))}\,, \qquad \textit{where} \quad p(t) = \boldsymbol{P}[|X| > t]\,, \tag{50}$$

and we assume that $p(t) > 0$ *and* $1 - p(t) > 0 \;\; \forall t > 0$. *Then we have:*

$\lim_{t\uparrow\infty} \boldsymbol{RAV}[\sigma_t^2] = \infty$;
$\lim_{t\downarrow 0} \boldsymbol{RAV}[\sigma_t^2] = 0$ *if the distribution of* X *has no atom at the origin:* $\boldsymbol{P}[X=0]=0$.

We abbreviate $\bar{p}(t) = 1 - p(t)$ in what follows.

$$\begin{aligned}
\boldsymbol{RAV}[\eta_t] \;&=\; \boldsymbol{E}\left[\eta_t(X)^2 X^2\right] \\
&=\; \frac{1}{p(t)\bar{p}(t)}\, \boldsymbol{E}\left[\left(1_{[|X|>t]} - p(t)\right)^2 X^2\right] \\
&=\; \frac{1}{p(t)\bar{p}(t)}\, \boldsymbol{E}\left[\left(1_{[|X|>t]} - 2\cdot 1_{[|X|>t]}\, p(t) + p(t)^2\right) X^2\right] \\
&=\; \frac{1}{p(t)\bar{p}(t)}\, \boldsymbol{E}\left[\left(1_{[|X|>t]}(1 - 2\,p(t)) + p(t)^2\right) X^2\right] \\
&=\; \frac{1}{p(t)\bar{p}(t)}\, \left(\boldsymbol{E}\left[1_{[|X|>t]} X^2\right](1 - 2\,p(t)) + p(t)^2\right) \\
&\geq\; \frac{1}{p(t)\bar{p}(t)}\, \left(p(t)\, t^2\,(1 - 2\,p(t)) + p(t)^2\right) \qquad \text{for} \quad p(t) \leq \frac{1}{2} \\
&=\; \frac{1}{\bar{p}(t)}\, \left(t^2\,(1 - 2\,p(t)) + p(t)\right) \\
&\geq\; t^2\,(1 - 2\,p(t)) + p(t) \\
&\sim\; t^2 \qquad \text{as} \qquad t \uparrow \infty.
\end{aligned}$$

For the following we note $1_{[|X|>t]} - p(t) = -1_{[|X|\le t]} + \bar{p}(t)$:

$$\begin{aligned}
\boldsymbol{RAV}[\eta_t] &= \boldsymbol{E}\left[\eta_t(X)^2 X^2\right] \\
&= \frac{1}{p(t)\bar{p}(t)}\,\boldsymbol{E}\left[\left(1_{[|X|\le t]} - \bar{p}(t)\right)^2 X^2\right] \\
&= \frac{1}{p(t)\bar{p}(t)}\,\boldsymbol{E}\left[\left(1_{[|X|\le t]} - 2\cdot 1_{[|X|\le t]}\,\bar{p}(t) + \bar{p}(t)^2\right) X^2\right] \\
&= \frac{1}{p(t)\bar{p}(t)}\,\boldsymbol{E}\left[\left(1_{[|X|\le t]}(1 - 2\,\bar{p}(t)) + \bar{p}(t)^2\right) X^2\right] \\
&= \frac{1}{p(t)\bar{p}(t)}\,\left(\boldsymbol{E}\left[1_{[|X|\le t]} X^2 (1 - 2\,\bar{p}(t))\right] + \bar{p}(t)^2\right) \\
&\le \frac{1}{p(t)\bar{p}(t)}\,\left(\bar{p}(t)\,t^2\,(1 - 2\,\bar{p}(t)) + \bar{p}(t)^2\right) \qquad \text{for} \quad \bar{p}(t) \le \frac{1}{2} \\
&= \frac{1}{p(t)}\,\left(t^2\,(1 - 2\,\bar{p}(t)) + \bar{p}(t)\right) \\
&\sim t^2 + \bar{p}(t) \qquad \text{as} \qquad t \downarrow 0,
\end{aligned}$$

assuming $\bar{p}(0) = P[X = 0] = 0$.

B.5 Details for Figure 5

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 5 the base function is, up to scale, as follows:

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

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

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

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

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

and hence

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

Figure 5 shows the functions as follows: $f(x)^2/\boldsymbol{E}[f^2(X)] = f(x)^2/s_2$.

B.6 Proof of Asymptotic Normality of $R\hat{A}V_j$, Section 9

We recall (45) for reference in the following form:

$$R\hat{A}V_j \;=\; \frac{\frac{1}{N}\langle (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\frac{1}{N}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\;\frac{1}{N}\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,. \tag{51}$$

For the denominators it is easy to show that

$$\begin{aligned} \tfrac{1}{N}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 &\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[\,\delta^2\,], \\ \tfrac{1}{N}\|\boldsymbol{X}_{j\hat{\bullet}}\|^2 &\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[\,X_j{}_\bullet{}^2\,]. \end{aligned} \tag{52}$$

For the numerator a CLT holds based on

$$\tfrac{1}{N^{1/2}}\langle (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle \;=\; \tfrac{1}{N^{1/2}}\langle (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^2, \boldsymbol{X}_{j\bullet}{}^2\rangle + O_P(N^{-1/2}). \tag{53}$$

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 $\{(Y_i, \vec{\boldsymbol{X}}_i)\}_{i=1\ldots N}$ to be i.i.d. sampled pairs, hence $\{(\delta_i^2, X_{i,j}{}_\bullet{}^2)\}_{i=1\ldots N}$ are N i.i.d. sampled pairs as well. Using the denominator terms (52) and Slutsky's theorem, we arrive at the first version of the CLT for $R\hat{A}V_j$:

$$N^{1/2}\,(R\hat{A}V_j - 1) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0,\, \frac{\boldsymbol{E}[\,\delta^4]}{\boldsymbol{E}[\,\delta^2]^2}\,\frac{\boldsymbol{E}[X_j{}_\bullet{}^4]}{\boldsymbol{E}[X_j{}_\bullet{}^2]^2} - 1\right)$$

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 $R\hat{A}V_j$:

$$N^{1/2}\,(R\hat{A}V_j - 1) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0,\, 3\,\frac{\boldsymbol{E}[X_j{}_\bullet{}^4]}{\boldsymbol{E}[X_j{}_\bullet{}^2]^2} - 1\right).$$

Details for the numerator (53), using notation of Sections 7.1 and 7.2, in particular $\boldsymbol{X}_{j\bullet} = \boldsymbol{X}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j\bullet}$ and $\boldsymbol{X}_{j\hat{\bullet}} = \boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}}$:

$$\begin{aligned} \langle (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle &= \langle\, ((\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2,\, (\boldsymbol{X}_{j\bullet} - \boldsymbol{X}_{-j}(\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} - \boldsymbol{\beta}_{-j\bullet}))^2\,\rangle \\ &= \langle\, \boldsymbol{\delta}^2 + (\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2 - 2\,\boldsymbol{\delta}\,(\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})), \\ &\qquad \boldsymbol{X}_{j\bullet}{}^2 + (\boldsymbol{X}_{-j}(\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} - \hat{\boldsymbol{\beta}}_{-j\bullet}))^2 - 2\,\boldsymbol{X}_{j\bullet}(\boldsymbol{X}_{-j}(\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} - \boldsymbol{\beta}_{-j\bullet}))\,\rangle \\ &= \langle\, \boldsymbol{\delta}^2, \boldsymbol{X}_{j\bullet}{}^2\,\rangle + \ldots \end{aligned} \tag{54}$$

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 predictors $\vec{\boldsymbol{X}}_i$: 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 (54) 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:

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

The term in the paren 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 (52): It is sufficient to consider the first denominator term.

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

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

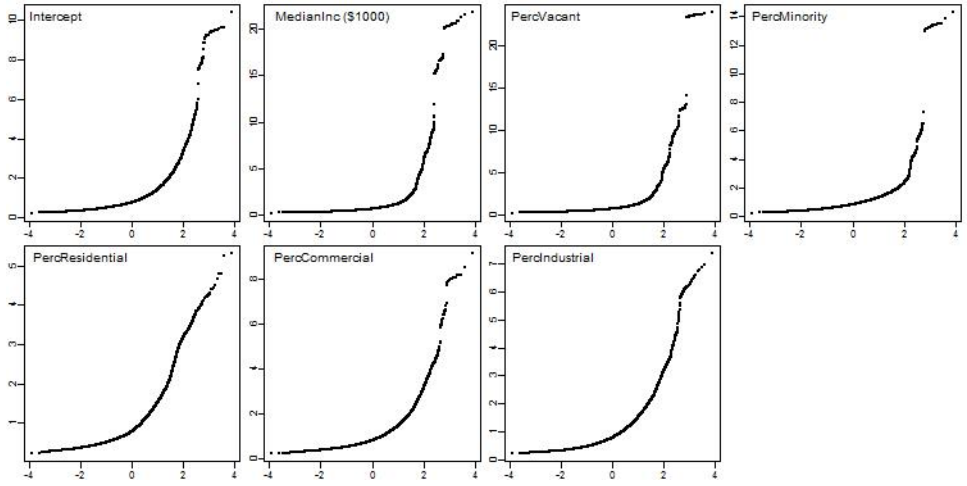

FIG 9. *Permutations distributions of* $\hat{\boldsymbol{RAV}}_j$ *for the LA Homeless Data*

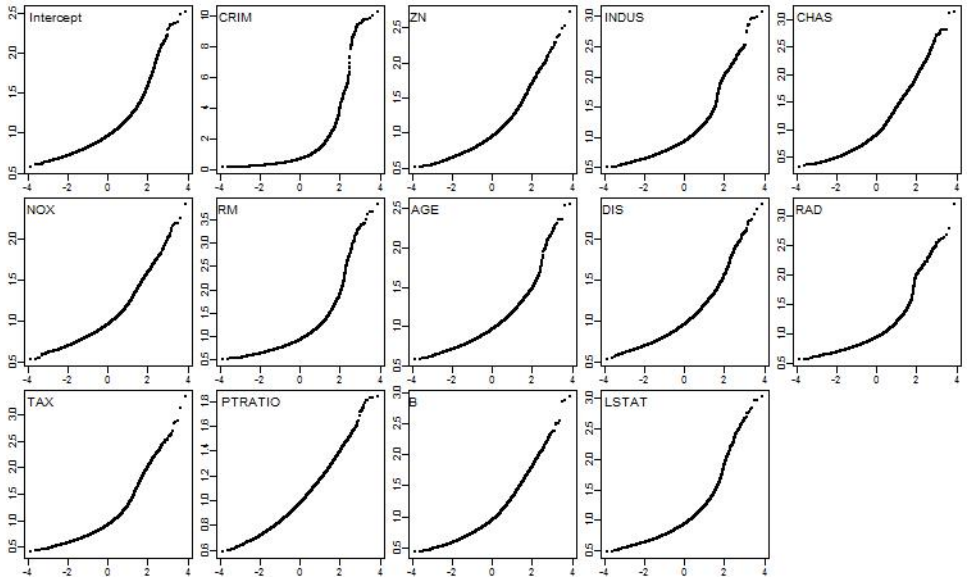

FIG 10. *Permutations distributions of* $\hat{\boldsymbol{RAV}}_j$ *for the Boston Housing Data*