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

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

Wharton – University of Pennsylvania[‡] and Amazon.com[§] and UNC at Chapel Hill[¶]

Dedicated to Halbert White (†2012)

Abstract.

We review and interpret aspects of Halbert White's early work which over thirty years ago inaugurated a "model-robust" form of statistical inference based on the "sandwich estimator" of standard error. This type of inference is asymptotically correct even under "model misspecification," that is, when the model is an approximation rather than a generative truth. Comparing the sandwich standard error to the usual but potentially incorrect standard errory, we show that asymptotically the discrepancy between the two can be of arbitrary magnitude either way. — Careful reading of White's work shows that the deepest consequences for inference arise from a synergistic effect (a "conspiracy") of nonlinearity and randomness of the regressors. This effect invalidates the ancillarity argument that justifies conditioning on the regressors when they are random: In the presence of nonlinearity, parameters do depend on the regressor distribution, and nonlinearity conspires with randomness of the regressors to generate a $1/\sqrt{N}$ contribution to sampling variability in the estimates.

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

Key words and phrases: Ancillarity of regressors, First and second order incorrect models, Model misspecification, Misspecification tests, Econometrics, Sandwich estimator of standard error, x-y bootstrap.

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

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

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

1. INTRODUCTION

Halbert White's basic sandwich estimator of standard error can be described as follows: In a linear model given by a regressor matrix $\boldsymbol{X}_{N\times(p+1)}$ and a response vector $\boldsymbol{y}_{N\times 1}$, start with the familiar derivation of the covariance matrix of the OLS coefficient estimate $\hat{\boldsymbol{\beta}}$, but allow heteroskedasticity, $\boldsymbol{V}[\boldsymbol{y}] = \boldsymbol{D}$ diagonal:

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

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

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

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

Lesser known is the following deeper result in one of White's (1980a, p. 162-3) less widely read articles: the sandwich estimator of standard error is asymptotically correct even in the presence of nonlinearity:

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

The term "heteroskedasticity-consistent" is an unfortunate choice as it obscures the fact that the same estimator of standard error is also "nonlinearity-consistent." Because of the relative obscurity of this important fact we will pay considerable attention to its implications. In particular we show how nonlinearity "conspires" with randomness of the regressors to make slopes dependent on the regressor distribution and to generate sampling variability all of their own even in the absence of noise (see Figures 2 and 4).

Side remarks:

- The term "nonlinearity" is meant here in the sense of (3), that is, first order misspecification of, or first order model deviation from, the linear model, $\boldsymbol{E}[\,\boldsymbol{y}\,|\boldsymbol{X}] - \boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{0}$. A different meaning of "nonlinearity", *not* intended here, occurs when the regressor matrix $\boldsymbol{X}$ contains multiple columns that are functions (polynomials, B-splines, ...) of an independent variable. We distinguish between "regressors" and "independent variables": Multiple regressors may be functions of the same independent variable.
- The sandwich estimator (2) is only the simplest version of its kind. Other versions were examined, for example, by MacKinnon and White (1985) and Long and Ervin (2000). Also, generalizations are pervasive in Generalized Estimating Equations (GEE; Liang and Zeger 1986; Diggle et al. 2002) and Generalized Method of Moments (GMM; Hansen 1982).

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

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] := (\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\sigma}^2, \qquad \hat{\sigma}^2 = \|\boldsymbol{r}\|^2/(N-p-1). \tag{4}$$

This yields finite-sample unbiased squared standard error estimators $\hat{\boldsymbol{SE}}^2_{lin}[\,\hat{\beta}_j\,] = \hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,]_{jj}$ if the model is first and second order correct: for some $\boldsymbol{\beta}$ and σ^2,

$$(5) \qquad \boldsymbol{E}[\,\boldsymbol{y}\,|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta} \ \text{ (linearity)}, \qquad \boldsymbol{V}[\,\boldsymbol{y}\,|\boldsymbol{X}] = \sigma^2\boldsymbol{I}_N \ \text{ (homoskedasticity)}.$$

Assuming also distributional correctness for the errors (normality), one obtains finite-sample correct tests and confidence intervals.

The analogous tests and confidence intervals based on the sandwich estimator have only an asymptotic justification, but their asymptotic validity holds under much weaker assumptions. In fact, it may rely on no more than the assumption that the rows $(\vec{\boldsymbol{x}}_i', y_i)$ of the data matrix $(\boldsymbol{X}, \boldsymbol{y})$ are i.i.d. samples from a joint multivariate distribution that have moments to some order. Thus sandwich-based theory provides asymptotically correct inference that is **assumption-lean** or **model-robust**; linear models theory provides then finite-sample correct inference that is **assumption-laden** or **model-dependent**. The question arises what sandwich-based inference is about: When no model is assumed, what are the parameters, and what is their meaning? A more radical question is: What sense does inference make when the model is wrong (Freedman 2006)?

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

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

A third goal of this article is to connect the assumption-lean framework to the "x-y bootstrap," which resamples observations $(\vec{\boldsymbol{x}}_i', y_i)$. In contrast, the "residual bootstrap" resamples residuals r_i. Theory exists to justify both types of bootstrap under different assumptions (see, for example, Freedman 1981, Mammen 1993). The x-y bootstrap can be asymptotically justified in the assumption-lean framework to produce standard error estimates that solve the same problem as the sandwich estimator. Indeed, a close connection exists: the sandwich estimator is the asymptotic limit of the M-of-N bootstrap when $M \to \infty$. Thus both may be called **assumption-lean** or **model-robust estimators** (Section 6).

A fourth goal of this article is to practically (Section 2) and theoretically (Section 9) compare the assumption-lean estimators with the linear models estimator. We define a ratio of asymptotic variances — "$\boldsymbol{RAV}$" for short — that describes the discrepancies between the two standard errors in the asymptotic limit. If there exists a discrepancy, $\boldsymbol{RAV} \neq 1$, it is assumption-lean estimators (sandwich or x-y bootstrap) that are asymptotically correct, and the usual standard error is then indeed asymptotically incorrect. It will be shown that the $\boldsymbol{RAV}$ can range from 0 to ∞ under certain scenarios, which gives insight into the nature of model deviations that affect standard errors.

A fifth goal is to estimate the $\boldsymbol{RAV}$ for use as a test statistic. We derive an asymptotic null distribution to test the presence of model violations that invalidate the usual standard error of a specific coefficient. Although the result can be called a "misspecification test," it is more usefully viewed as a discrepancy test for standard errors, separately for each coefficient (Section 10).

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

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

Throughout we use precise notation for clarity, yet this article is not very technical. The majority of results is elementary, not new, and stated without regularity conditions. The linear model is used to allow explicit calculations, but most conclusions generalize in some form to many other models. The linear model allows the clearest analysis of issues relating to regressor randomness and the effects of nonlinearity and heteroskedasticity. The emphasis is on detailed and concrete insights rather than novelty of technical results.

The article is written for readers who are not very familiar with the sandwich estimator. Readers who are may skim the article for appearances of the nonlinearity η, which is the aspect of this work that is least known. Readers may also prefer to selectively browse the tables and figures and then read associated sections that seem most germane.

Note on Terminology. We use the following interchangeably: misspecification = model deviation; assumption-laden = model-dependent = usual (standard error); assumption-lean = model-robust (standard error); 1st order = conditional mean (model specification); 2nd order = conditional variance (model specification).

2. DISCREPANCIES BETWEEN STANDARD ERRORS ILLUSTRATED

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

The ratios $\boldsymbol{SE}_{sand}/\boldsymbol{SE}_{boot}$ show that the sandwich and bootstrap estimators are in good agreement. Not so for the linear models estimates: we have $\boldsymbol{SE}_{boot}, \boldsymbol{SE}_{sand} > \boldsymbol{SE}_{lin}$ for the regressors `PercVacant`, `PercCommercial` and `PercIndustrial`, and $\boldsymbol{SE}_{boot}, \boldsymbol{SE}_{sand} < \boldsymbol{SE}_{lin}$ for `Intercept`, `MedianInc ($1000)`, `PercResidential`. Only for `PercMinority` is $\boldsymbol{SE}_{lin}$ off by less than 10% from $\boldsymbol{SE}_{boot}$ and $\boldsymbol{SE}_{sand}$. The

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

TABLE 1
LA Homeless Data: Comparison of Standard Errors.

discrepancies affect outcomes of some of the t-tests: Under linear models theory the regressors `PercCommercial` and `PercIndustrial` have commanding t-values of 2.700 and 2.818, respectively, which are reduced to unconvincing values below 1.9 and 1.6, respectively, if the x-y bootstrap or the sandwich estimator are used. On the other hand, for `MedianInc ($K)` the t-value -0.977 from linear models theory becomes borderline significant with the bootstrap or sandwich estimator if the plausible one-sided alternative with negative sign is used.

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

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

3. THE POPULATION FRAMEWORK

3.1 Targets of Estimation

To make standard errors meaningful it is necessary to first define targets of estimation. As mentioned in the introduction, parameters of generative models are reinterpreted as statistical functionals that are well-defined for a large nonparametric class of data distributions. In an assumption-lean population framework for linear regression with random regressors the ingredients are regressor random variables X_1, ..., X_p and a response random variable Y. For now the only assumption is that they have a joint distribution,

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

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

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

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

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

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

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

$$\mu(\vec{\boldsymbol{X}}) \;:=\; \mathrm{argmin}_{f(\vec{\boldsymbol{X}})\in L_2(\boldsymbol{P})} \boldsymbol{E}[(Y - f(\vec{\boldsymbol{X}}))^2] \;=\; \boldsymbol{E}[\,Y \,|\, \vec{\boldsymbol{X}}\,]\,. \tag{6}$$

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

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

$$\boldsymbol{\beta}(\boldsymbol{P}) := \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \boldsymbol{E}[(Y - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2] \qquad = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}Y\,] \tag{7}$$

$$= \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \boldsymbol{E}[(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2] = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})\,] \tag{8}$$

The right-most expressions in (7) and (8) follow from the normal equations:

$$\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\,\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}Y] \;=\; \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']\,\boldsymbol{\beta} - \boldsymbol{E}[\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})] \;=\; \boldsymbol{0}. \tag{9}$$

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

Generalizations:

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

$$\boldsymbol{\theta}(\boldsymbol{P}) = \mathrm{argmin}_{\boldsymbol{\theta}}\boldsymbol{E}_{\boldsymbol{P}}[-\log p(Y|\vec{\boldsymbol{X}};\boldsymbol{\theta})], \tag{10}$$

  or by solving the associated moment conditions/estimating equations,

$$\boldsymbol{E}_{\boldsymbol{P}}[\partial/\partial\boldsymbol{\theta}\, \log p(Y|\vec{\boldsymbol{X}};\boldsymbol{\theta})] = \boldsymbol{0}. \tag{11}$$

  Under mild regularity conditions we have $\boldsymbol{\theta}(\boldsymbol{P}) = \boldsymbol{\theta}_0$ if the actual conditional data distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ has density $p(y\,|\,\vec{\boldsymbol{x}};\boldsymbol{\theta}_0)$. The point is, however, that $\boldsymbol{\theta}(\boldsymbol{P})$ is defined for a large class of data distributions outside of the model $p(y\,|\,\vec{\boldsymbol{x}};\boldsymbol{\theta})$. The two-fold role of the model is 1) to provide a heuristic for a loss function $\mathcal{L}(\boldsymbol{\theta}; y, \vec{\boldsymbol{x}}) = -\log p(y\,|\,\vec{\boldsymbol{x}};\boldsymbol{\theta})$, and 2) to act as an approximation to the actual conditional data distribution $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. An early adopter of this point of view is Kent (1982).
- Another generalization that overlaps with the previous is the method of moments (MM) where a moment condition $\boldsymbol{E}_{\boldsymbol{P}}[\boldsymbol{\psi}(Y,\vec{\boldsymbol{X}};\boldsymbol{\theta})] = \boldsymbol{0}$ defines a statistical functional $\boldsymbol{\theta}(\boldsymbol{P})$. This condition is no longer required to be the stationarity condition of any optimization, in particular it is not necessarily the score function of a likelihood. A seminal work that inaugurated asymptotic theory for very general moment conditions is by Huber (1967). For OLS, $\boldsymbol{\psi}(y,\vec{\boldsymbol{x}};\boldsymbol{\beta}) = \vec{\boldsymbol{x}}\vec{\boldsymbol{x}}'\,\boldsymbol{\beta} - \vec{\boldsymbol{x}}y$, so the moment condition specializes to (9).

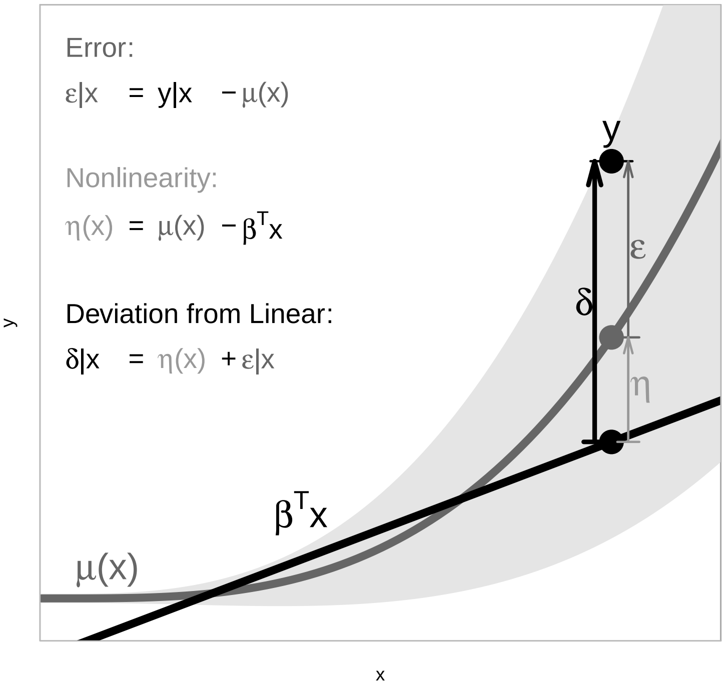

Fig 1. *Illustration of the decomposition* (12).

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

3.2 The Canonical Noise-Nonlinearity Decomposition

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

$$\begin{aligned} Y &= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \underbrace{(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\vec{\boldsymbol{X}})} + \underbrace{(Y - \mu(\vec{\boldsymbol{X}}))} \\ &= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \underbrace{\eta(\vec{\boldsymbol{X}}) \quad + \quad \epsilon} \\ &= \boldsymbol{\beta}'\vec{\boldsymbol{X}} + \qquad \delta \end{aligned} \tag{12}$$

We call ϵ the noise and η the nonlinearity, while for δ there is no standard term, but "population residual" may suffice; see Table 2. The following list contains mutual relations between the regressors and the components of the canonical decompositions, as well as some further definitions:

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

TABLE 2
Random variables and their canonical decompositions.

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

$$\boldsymbol{E}[\,\epsilon\, f(\vec{\boldsymbol{X}})\,] \;=\; 0 \quad \forall f(\vec{\boldsymbol{X}}) \in L_2(\boldsymbol{P}_{\vec{\boldsymbol{X}}}), \tag{13}$$

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

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

$$\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\eta\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\epsilon\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\delta\,] \;=\; \boldsymbol{0}. \tag{14}$$

  The first, $\eta \perp \vec{\boldsymbol{X}}$, holds because by (8) η is the population residual of the OLS linear regression of $\mu(\vec{\boldsymbol{X}})$ on $\vec{\boldsymbol{X}}$; the second, $\epsilon \perp \vec{\boldsymbol{X}}$, follows from (13); finally, $\delta \perp \vec{\boldsymbol{X}}$ because $\delta = \eta + \epsilon$.

- ***Marginal centering***, unconditional, is a special case of (14) due to the inclusion of an intercept in $\vec{\boldsymbol{X}}$:

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

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

$$\sigma^2(\vec{\boldsymbol{X}}) := \boldsymbol{V}[\,\epsilon\,|\vec{\boldsymbol{X}}] \;=\; \boldsymbol{E}[\epsilon^2\,|\,\vec{\boldsymbol{X}}] \;\overset{\boldsymbol{P}}{<}\; \infty. \tag{16}$$

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

$$m^2(\vec{\boldsymbol{X}}) \;:=\; \boldsymbol{E}[\,\delta^2\,|\,\vec{\boldsymbol{X}}] \;=\; \eta^2(\vec{\boldsymbol{X}}) + \sigma^2(\vec{\boldsymbol{X}}). \tag{17}$$

  The decomposition follows from $\delta = \eta + \epsilon$ and $\epsilon \perp \eta(\vec{\boldsymbol{X}})$ due to (13).

- ***Mean squared functionals***:

$$\begin{array}{rcll} \eta^2(\boldsymbol{P}) & := & \boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})\,], & \text{mean squared nonlinearity,} \\ \sigma^2(\boldsymbol{P}) & := & \boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\,] = \boldsymbol{E}[\,\epsilon^2\,], & \text{mean noise variance,} \\ m^2(\boldsymbol{P}) & := & \boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,], & \text{mean or plain MSE.} \end{array} \tag{18}$$

  All expectations, except for $\boldsymbol{E}[\,\epsilon^2\,]$, are w.r.t. $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. From (17) follows

$$m^2(\boldsymbol{P}) = \eta^2(\boldsymbol{P}) + \sigma^2(\boldsymbol{P}). \tag{19}$$

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

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

3.3 Error Terms and Random Regressors: Uncorrelated versus Independent

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

- ***Weak-sense error terms***: e and $\vec{\boldsymbol{X}}$ are orthogonal, $\boldsymbol{E}[e\,\vec{\boldsymbol{X}}] = \boldsymbol{0}$.
- ***Medium-sense error terms***: e and $L_2(\vec{\boldsymbol{X}})$ are orthogonal.
- ***Strong-sense error terms***: e and $\vec{\boldsymbol{X}}$ are independent.

Error terms in the weak sense permit $e = \delta = \eta + \epsilon$ by weak-sense orthogonality of δ w.r.t. $\vec{\boldsymbol{X}}$ (14), hence they may include heteroskedastic noise as well as nonlinearity, so that misspecification to first or second order is meaningless. A medium sense error term allows heteroskedastic $e = \epsilon$, but requires $\eta \overset{\boldsymbol{P}}{=} 0$, so that misspecification to first order is meaningful but not to second order. Error terms in the strong sense must be homoskedastic and nonlinearity-free, so that the notion of misspecification to first and second order is meaningful.

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

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

4. NON-ANCILLARITY OF THE REGRESSOR DISTRIBUTION

4.1 The Breakdown of the Ancillarity Argument

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

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

referring to the model densities of $\boldsymbol{P}_{\vec{\boldsymbol{X}},Y}$, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, respectively. The parameter of interest is $\boldsymbol{\theta}$ while the regressor density $p(\vec{\boldsymbol{x}})$ acts as a "nonparametric nuisance parameter." Ancillarity of $p(\vec{\boldsymbol{x}})$ in relation to $\boldsymbol{\theta}$ is immediately recognized by forming likelihood ratios $p(y, \vec{\boldsymbol{x}}; \boldsymbol{\theta}_1)/p(y, \vec{\boldsymbol{x}}; \boldsymbol{\theta}_2) = p(y \,|\, \vec{\boldsymbol{x}}; \boldsymbol{\theta}_1)/p(y \,|\, \vec{\boldsymbol{x}}; \boldsymbol{\theta}_2)$ which are free of $p(\vec{\boldsymbol{x}})$. (For a fuller definition of ancillarity see Appendix B.) This logic

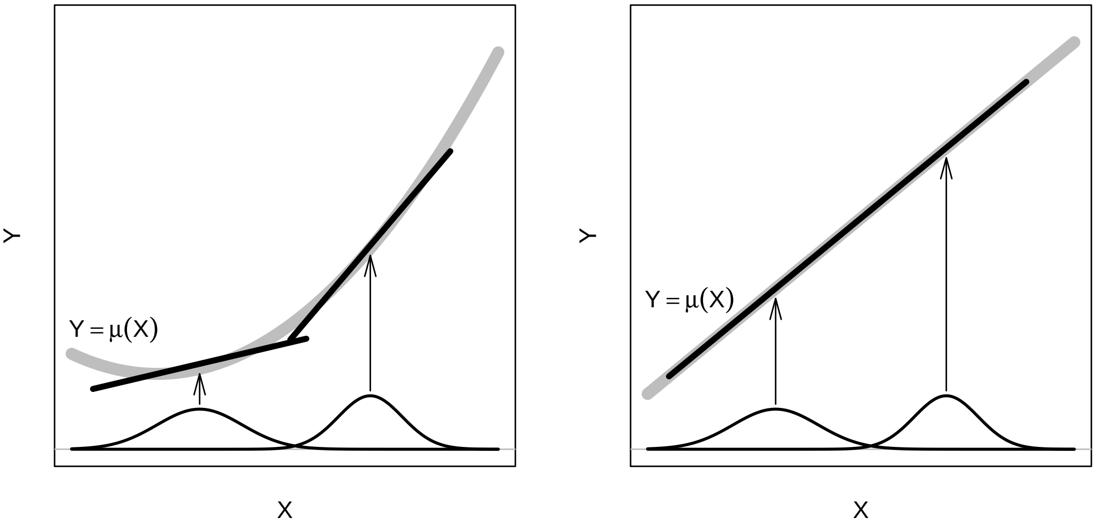

Fig 2. *Illustration of the dependence of the population OLS solution on the marginal distribution of the regressors: The left figure shows dependence in the presence of nonlinearity; the right figure shows independence in the presence of linearity.*

is valid if the conditional model $p(y \,|\, \vec{\boldsymbol{x}};\, \boldsymbol{\theta})$ is correct. The following proposition describes for linear models the ways in which ancillarity is broken if the model is an approximation rather than a truth.

Proposition 4.1:

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

(These are loose statements; see Appendix C.2 for more precision.) The first part of the proposition is best explained with a graphical illustration: Figure 2 shows single regressor situations with a nonlinear and a linear mean function, respectively, and the same two regressor distributions. The two population OLS lines for the two regressor distributions differ in the nonlinear case and they are identical in the linear case. (This observation appears first in White (1980a, p. 155f); to see the correspondence, identify Y with his $g(Z) + \epsilon$.)

Ancillarity of regressors is sometimes informally explained as the regressor distribution being independent of, or unaffected by, the parameters of interest. This phrasing has things upside down: It is not the parameters that affect the regressor distribution; it is the regressor distribution that affects the parameters.

4.2 Implications of the Dependence of Slopes on Regressor Distributions

A first practical implication, illustrated by Figure 2, is that two empirical studies that use the same regressors, the same response variable, and the same model, may yet estimate different parameter values, $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$. What may

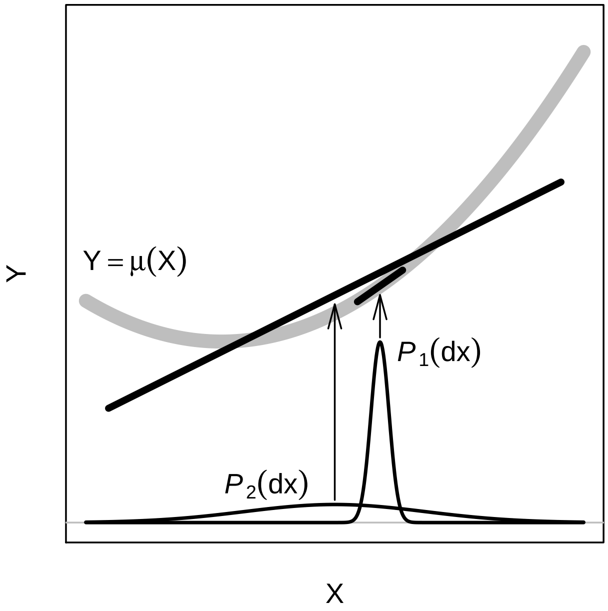

FIG 3. *Illustration of the interplay between regressors' high-density range and nonlinearity: Over the small range of $\boldsymbol{P}_1$ the nonlinearity will be undetectable and immaterial for realistic sample sizes, whereas over the extended range of $\boldsymbol{P}_2$ the nonlinearity is more likely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

seem to be superficially contradictory inferences from the two studies may be compatible if 1) the true response surface $\mu(\vec{\boldsymbol{x}})$ is not linear and 2) the regressors' high-density regions differ between studies. Differences in regressor distributions can become increasingly complex for larger regressor dimensions or, worse, as $p \to \infty$. Differences in estimated parameter values often become visible in meta-analyses and may be interpreted as "parameter heterogeneity." The source of this heterogeneity may be differences in covariate distributions combined with nonlinearities relative to the fitted model.

A second practical implication, illustrated by Figure 3, is that misspecification is a function of the regressor range: Over a narrow range a model has a better chance of appearing "well-specified" because approximations work better over narrow ranges. In the figure the narrow range of the regressor distribution $\boldsymbol{P}_1(d\vec{\boldsymbol{x}})$ is the reason why the linear approximation is excellent, hence the model very nearly "well-specified," whereas the wide range of $\boldsymbol{P}_2(d\vec{\boldsymbol{x}})$ is the reason for the gross "misspecification" of the linear approximation. This is a general issue that holds even in the most successful theories, those of physics, which at this point in history have limited ranges of validity as well.

5. OBSERVATIONAL DATASETS, ESTIMATION, AND CLTS

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

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

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

to its two sources, noise and nonlinearity, with emphasis on the latter.

5.1 Observational Datasets and Estimation

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

$$\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}, \qquad \boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\eta} + \boldsymbol{\epsilon}. \tag{20}$$

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

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

even though the associated random variables are orthogonal to X_j in the population: $\boldsymbol{E}[\,\delta X_j] = 0$, $\boldsymbol{E}[\,\epsilon X_j] = 0$, $\boldsymbol{E}[\,\eta(\vec{\boldsymbol{X}}) X_j] = 0$, according to (14).

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

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\tilde{\boldsymbol{\beta}}} \, \| \boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} \|^2 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \, . \tag{21}$$

Because we are not conditioning on $\boldsymbol{X}$, randomness of $\hat{\boldsymbol{\beta}}$ stems from $\boldsymbol{Y}$ as well as $\boldsymbol{X}$. The sample residual vector $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, which arises from $\hat{\boldsymbol{\beta}}$, is distinct from the population residual vector $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, which arises from $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$. If we write $\hat{\boldsymbol{P}}$ for the empirical distribution of the N observations, then $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}})$ is the plug-in estimate.

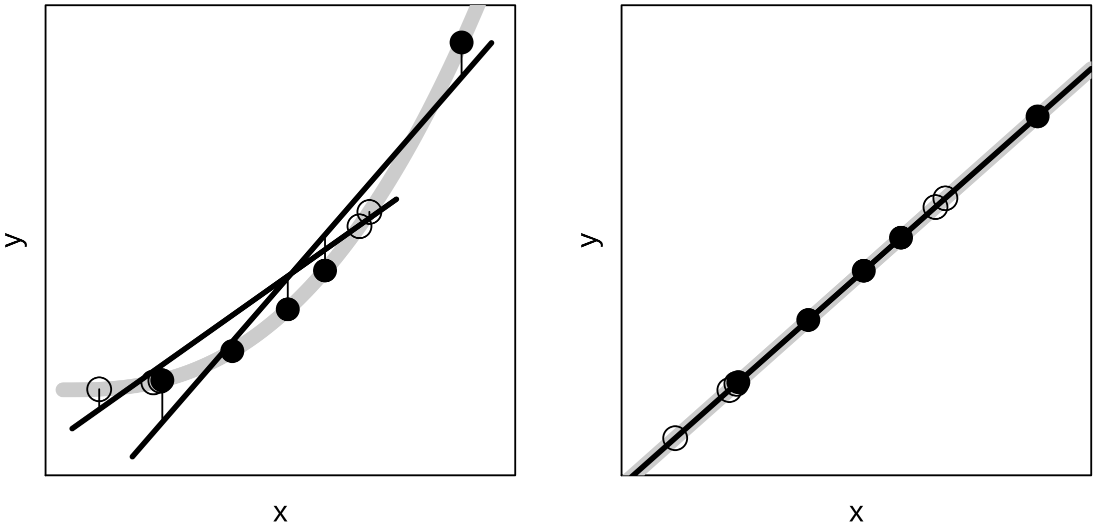

Fig 4. *Noise-less Response: The filled and the open circles represent two "datasets" from the same population. The x-values are random; the y-values are a deterministic function of x: $y = \mu(x)$ (shown in gray).*
Left: The true response $\mu(x)$ is nonlinear; the open and the filled circles have different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

5.2 Decomposition of OLS Estimates According to Noise and Nonlinearity

In linear models theory, as in any fixed-$\boldsymbol{X}$ theory, 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 OLS solution $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$. Hence fixed-$\boldsymbol{X}$ theory misses out on a term $\boldsymbol{E}[\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}] - \boldsymbol{\beta}$ which will be seen to contribute in the order of $1/\sqrt{N}$ to the unconditional standard error of $\hat{\boldsymbol{\beta}}$ in the presence of a nonlinearity, $\boldsymbol{E}[\eta^2] > 0$.

The random vector $\boldsymbol{E}[\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]$ is naturally placed between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; (\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]) \;+\; (\boldsymbol{E}[\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}] - \boldsymbol{\beta}). \tag{22}$$

This decomposition corresponds to the canonical noise-nonlinearity decomposition $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$:

Definition and Lemma: *Define "Estimation Offsets" or "EOs" as follows:*

$$\boxed{\begin{array}{lcccl} \textit{Total EO} & := & \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} & = & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\delta}, \\ \textit{Noise EO} & := & \hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}] & = & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}, \\ \textit{Nonlinearity EO} & := & \boldsymbol{E}[\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}] - \boldsymbol{\beta} & = & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}. \end{array}} \tag{23}$$

The right hand equalities follow from the decompositions (20), $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, and these facts:

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

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

Generalizations: The three *EO*s of the above Lemma can be defined for quite arbitrary estimators $\hat{\boldsymbol{\theta}}$ in regressor-response data, and this would be the starting point of a more general comparative analysis of fixed-$\boldsymbol{X}$ versus random-$\boldsymbol{X}$ regression. We work out the details for OLS estimators for their clean separation of first and second order properties and the ensuing crispness of tracing the *EO* $\hat{\boldsymbol{\theta}} - \boldsymbol{E}[\,\hat{\boldsymbol{\theta}}\,|\boldsymbol{X}]$ to noise and the *EO* $\boldsymbol{E}[\,\hat{\boldsymbol{\theta}}\,|\boldsymbol{X}] - \boldsymbol{\theta}$ to nonlinearity.

5.3 Random X and Nonlinearity as a Source of Sampling Variation

Linear models theory is about sampling variability due to noise represented by the noise EO. Because of its assumption of well-specification it ignores the other source of sampling variability, nonlinearity in the presence of random regressors as represented by the nonlinearity EO. This latter source is best illustrated in a noise-free situation: Consider a response that is a deterministic but nonlinear function of the regressors, $Y = \mu(\vec{\boldsymbol{X}})$, so that $\boldsymbol{\epsilon} = \mathbf{0}$ but $\boldsymbol{\eta} \neq \mathbf{0}$. There exists sampling variability in $\hat{\boldsymbol{\beta}}$ due to the nonlinearity $\boldsymbol{\eta}$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}$, in conjunction with the randomness of the regressors — the “conspiracy” in the title of this article. (Noise-free nonlinear responses occur in practice when outputs from expensive deterministic simulation experiments are modeled based on inputs.)

Figure 4 illustrates the situation with a single-regressor example by showing the OLS lines fitted to two “datasets” consisting of $N = 5$ regressor values each. The random differences between datasets cause the fitted line to exhibit sampling variability under nonlinearity (left hand figure), which is absent under linearity (right hand figure). Comparing this figure with the earlier Figure 2, we see that the effects illustrated in both are the same, but Figure 2 shows it for different populations while Figure 4 shows it for different datasets. Thus nonlinearity creates complications on two levels: (1) in the definition of the population OLS parameter, which becomes dependent on the regressor distribution, and (2) through the creation of sampling variability in $\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}]$ 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 problem with linear models theory in this situation is that it confuses nonlinearity with noise. The consequences of this confusion for statistical inference will be examined in Section 9.4. They seep into the residual bootstrap which assumes the residuals to originate from exchangeable noise. By comparison, the sandwich estimator and the x-y bootstrap get statistical inference right even in the noise-free nonlinear case, at least asymptotically. The justification derives from central limit theorems which are described next.

5.4 Assumption-Lean Central Limit Theorems

The proofs of the following are standard; see Appendix C.3.

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

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

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

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

Note that the contribution of the nonlinearity in combination with the randomness of the regressors is of the same order $1/\sqrt{N}$ as the contribution of the noise. The CLTs are shown in terms of the decomposition (17), $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$, but by (16,17) $m^2(\vec{\boldsymbol{X}})$ can be replaced by δ^2 and $\sigma^2(\vec{\boldsymbol{X}})$ by ϵ^2:

$$\boldsymbol{E}[\, m^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] \;=\; \boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'], \qquad \boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] \;=\; \boldsymbol{E}[\,\epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']. \tag{24}$$

Consider some special cases:

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

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

  The sandwich form is solely due to heteroskedasticity.

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

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

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

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

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

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

Generalizations: The CLT for $\hat{\boldsymbol{\beta}}$ is a very special case of assumption-lean CLTs for moment conditions that have been known at least since Huber (1967). Assuming a generic vector moment condition $\boldsymbol{E_P}[\boldsymbol{\psi}(Y, \vec{\boldsymbol{X}}; \boldsymbol{\theta})] = \mathbf{0}$ that defines a statistical functional $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{P})$ for $(Y, \vec{\boldsymbol{X}}) \sim \boldsymbol{P}$ (Subsection 3.1), and estimating $\boldsymbol{\theta}(\boldsymbol{P})$ from i.i.d. samples through plug-in, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{P}_N)$, there holds under technical conditions the following CLT, where $\boldsymbol{\Lambda}(\boldsymbol{\theta}) := \partial_{\boldsymbol{\theta}}\boldsymbol{E}[\boldsymbol{\psi}(Y, \vec{\boldsymbol{X}}; \boldsymbol{\theta})]$:

$$\sqrt{N}\,\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\Big(\mathbf{0},\, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\boldsymbol{V}[\boldsymbol{\psi}(Y, \vec{\boldsymbol{X}}; \boldsymbol{\theta})]\,\boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}\Big)\,. \tag{25}$$

This specializes to an assumption-lean CLT for ML estimation where $\boldsymbol{\psi}(y, \vec{\boldsymbol{x}}; \boldsymbol{\theta}) = -\partial_{\boldsymbol{\theta}} \log p(y|\vec{\boldsymbol{x}}; \boldsymbol{\theta})$. The assumption-laden CLT is obtained by assuming that $\boldsymbol{P}$ is given by a model density $p(y|\vec{\boldsymbol{x}}; \boldsymbol{\theta})$ for some unknown $\boldsymbol{\theta}$, in which case $\boldsymbol{V}[\boldsymbol{\psi}(Y, \vec{\boldsymbol{X}}; \boldsymbol{\theta})] = \boldsymbol{\Lambda}(\boldsymbol{\theta})$, so that the asymptotic normal distribution becomes $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1}\right)$.

It would be possible to produce separate CLTs for the *EO*s $\hat{\boldsymbol{\theta}} - \boldsymbol{E}[\,\hat{\boldsymbol{\theta}}\,|\boldsymbol{X}]$ and $\boldsymbol{E}[\,\hat{\boldsymbol{\theta}}\,|\boldsymbol{X}] - \boldsymbol{\theta}$ for general MM estimators.

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

Empirically one observes that standard error estimates obtained from the x-y bootstrap and from the sandwich estimator are generally close to each other. This is intuitively unsurprising as they both estimate the same asymptotic variance, that of the first CLT in Proposition 5.4. A closer connection between them will be established below.

6.1 The Plug-In Sandwich Estimator of Asymptotic Variance

According to Proposition 5.4 and (24) the asymptotic variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\boldsymbol{AV}[\hat{\boldsymbol{\beta}}] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\,\boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\;\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}. \tag{26}$$

The sandwich estimator is then the plug-in version of (26) where δ^2 is replaced by residuals and population expectations $\boldsymbol{E}[...]$ by sample means $\hat{\boldsymbol{E}}[...]$:

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

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

$$\begin{aligned} \hat{\boldsymbol{AV}}_{sand} \;&:=\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\,\hat{\boldsymbol{E}}[\,r^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\;\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1} \\ &=\; N\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\,(\boldsymbol{X}'\boldsymbol{D}(\boldsymbol{r})^2\,\boldsymbol{X})\,(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned} \tag{27}$$

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

$$\hat{\boldsymbol{SE}}_{sand}[\hat{\beta}_j] \;:=\; \tfrac{1}{N^{1/2}}\,(\hat{\boldsymbol{AV}}_{sand})_{jj}^{1/2}. \tag{28}$$

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 (28) would seem to be sensible in analogy to the linear models estimator ("HC1" ibid.). More detailed modifications have been proposed whereby individual residuals are corrected for their reduced conditional variance according to $\boldsymbol{V}[r_i|\boldsymbol{X}] = \sigma^2(1 - H_{ii})$ under homoskedasticity and ignoring nonlinearity ("HC2" ibid.). Further modifications include a version based on the jackknife ("HC3" ibid.) using leave-one-out residuals. An obvious alternative is estimating asymptotic variance with the x-y bootstrap, to which we now turn.

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

To connect the sandwich estimator to the bootstrap we need the M-of-N bootstrap whereby the *resample size* M is allowed to differ from the sample size N. It is 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 $M < N$. The M-of-N bootstrap for $M \ll N$ "works" more often than the conventional N-of-N bootstrap; see Bickel, Götze and van Zwet (1997) who showed that the favorable properties of $M \ll N$ subsampling obtained by Politis and Romano (1994) carry over to the $M \ll N$ bootstrap. Because we are here concerned only with well behaved OLS estimation, there is no reason to resort to $M \ll N$; instead, we consider bootstrap resampling for the extreme case $M \gg N$, namely, the limit $M \to \infty$.

The crucial observation is as follows: Because resampling is i.i.d. sampling from some distribution, there holds a CLT as the resample size grows, $M \to \infty$. It is immaterial that, in this case, the sampled distribution is the empirical distribution $\boldsymbol{P}_N$ of a given dataset $\{(Y_i, \vec{\boldsymbol{X}}_i)\}_{i=1...N}$, which is frozen of size N as $M \to \infty$.

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

$$(29)\quad M^{1/2}\,(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\Big(\mathbf{0},\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\,\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\;\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\Big).$$

This is a straight application of the CLT of the previous section to the empirical distribution of the data, where the middle part (the “meat”) of the asymptotic formula is based on the empirical counterpart $r_i^2 = (Y_i – \vec{\boldsymbol{X}}_i{}'\hat{\boldsymbol{\beta}})^2$ of $\delta^2 = (Y – \vec{\boldsymbol{X}}'\boldsymbol{\beta})^2$. A comparison of (27) and (29) results in the following:

Corollary 6.2: *The sandwich estimator* (27) *is the asymptotic variance estimated by the M-of-N bootstrap in the limit $M \to \infty$ for a fixed sample of size N.*

The sandwich estimator has the advantage that it results in unique standard error values whereas bootstrap standard errors have simulation error in practice. On the other hand, the x-y bootstrap is more flexible because the bootstrap distribution can be used to generate confidence intervals that are second order correct (see, e.g., Efron and Tibshirani 1994; Hall 1992).

Further connections are mentioned by MacKinnon and White (1985): Some forms of the sandwich estimator were independently derived by Efron (1982, p. 18f) using the infinitesimal jackknife, and by Hinkley (1977) using a “weighted jackknife.” See Weber (1986) for a concise comparison in the 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).

Generalizations: Sandwich estimators of standard error exist for a large class of well-behaved MM parameter estimators. They are obtained by plug-in into the asymptotic variance given by their CLTs (25):

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

It can again be shown that these sandwich estimators are the limits of the M-of-N bootstrap when $M \to \infty$ and N is fixed. Squared standard error estimates of parameter estimates are found in the diagonal of $\hat{\boldsymbol{AV}}/n$.

7. ADJUSTED REGRESSORS

The following adjustment formulas are standard but will be stated explicitly for their importance. They express the slopes of multiple regressions as slopes of simple regressions using adjusted single regressors. Subsequently the formulas will be used for the interpretation of regression slopes in the presence of nonlinearity (Section 8), the analysis of discrepancies between asymptotically proper and improper standard errors (Section 9), and a test of discrepancy between the two (Section 10).

7.1 Adjustment in Populations

Define the population-adjusted regressor random variable $X_{j\bullet}$ to be the "residual" of the population regression of X_j, used as the response, on all other regressors. Similarly, Y can be population-adjusted for all regressors other than X_j. To this end collect all other regressors in the random p-vector $\vec{\boldsymbol{X}}_{\text{-}j} = (1, X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)'$, and let

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

The response Y can be adjusted similarly, and we may denote it by $Y_{\bullet\text{-}j}$ to indicate that X_j is not among the adjustors, which is implicit in the adjustment of X_j. The simple regression through the origin of Y or $Y_{\bullet\text{-}j}$ on $X_{j\bullet}$ yields the multiple regression coefficient $\beta_j = \beta_j(\boldsymbol{P})$ of the population regression of Y on $\vec{\boldsymbol{X}}$:

$$(30) \qquad \beta_j \;=\; \frac{\boldsymbol{E}[\,Y_{\bullet\text{-}j}X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \;=\; \frac{\boldsymbol{E}[\,Y X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \;=\; \frac{\boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}})X_{j\bullet}]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]}\,.$$

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

7.2 Adjustment in Samples

Define the sample-adjusted regressor column $\boldsymbol{X}_{j\hat{\bullet}}$ to be the residual vector of the sample regression of $\boldsymbol{X}_j$, used as the response vector, on all other regressors. Similarly, $\boldsymbol{Y}$ can be sample-adjusted for all regressors other than $\boldsymbol{X}_j$. To this end collect all regressor columns other than $\boldsymbol{X}_j$ in a $N \times p$ random regressor matrix $\boldsymbol{X}_{\text{-}j} = [\boldsymbol{1}, ..., \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, ...]$ and let

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

(Note the use of hat notation "$\hat{\bullet}$" to distinguish it from population-based adjustment "$\bullet$".) The response vector $\boldsymbol{Y}$ can be sample-adjusted similarly, and we may denote it by $\boldsymbol{Y}_{\hat{\bullet}\text{-}j}$ to indicate that $\boldsymbol{X}_j$ is not among the adjustors. Finally, the simple regression through the origin of $\boldsymbol{Y}$ or $\boldsymbol{Y}_{\hat{\bullet}\text{-}j}$ on $\boldsymbol{X}_{j\bullet}$ yields the coefficient estimate $\hat{\beta}_j$ of the multiple regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$:

$$(31) \qquad \hat{\beta}_j \;=\; \frac{\langle \boldsymbol{Y}_{\hat{\bullet}\text{-}j}, \boldsymbol{X}_{j\hat{\bullet}} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2} \;=\; \frac{\langle \boldsymbol{Y}, \boldsymbol{X}_{j\hat{\bullet}} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,.$$

Generalizations: The adjustment formalism seems peculiar to OLS. It holds, however, with caveats also for generalized linear models when interpreted as iteratively reweighted LS problems. In this case, weighted adjustment formulas hold with the data-driven weights at convergence of the iteratively reweighted OLS iterations. A similar comment holds for M-estimation in robust regression.

8. THE MEANING OF SLOPES IN THE PRESENCE OF NONLINEARITY

A first use of regressor adjustment is for proposing a meaning of linear slopes in the presence of nonlinearity, and thereby responding to Freedman's (2006, p. 302) objection: "... it is quite another thing to ignore bias [nonlinearity]. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter." Against this view one may hold that the parameter is not intrinsically wrong, rather, it is in need of a useful interpretation. Intuitively, a linear fit gives a sense of the direction, up or down, of association between a regressor and the response in the presence of other regressors. (One may agree with Freedman if the sole goal is response prediction.)

The issue is that, in the presence of nonlinearities, slopes lose their common interpretation: β_j is no longer the average difference in Y associated with a unit difference in X_j at fixed levels of all other X_k. The challenge is to provide an alternative interpretation that is valid and intuitive even in the presence of nonlinearities. As mentioned, a plausible approach is to work with adjusted variables so that it is sufficient to solve the problem for the case of simple regressions through the origin. Regression slopes can then be interpreted as weighted averages of "case-wise" and "pairwise" slopes in a sense to be made precise. This interpretation will hold even for regressors that are nonlinearly related to each other, as in $X_2 = X_1^2$ and $X_3 = X_1 X_2$. The reason is that the clause "at fixed levels of all other regressors" will no longer be used and reference will be made instead to "(linearly) adjusted regressors," which is a notion that is meaningful even for nonlinearly associated regressors ("linearly" will be dropped in what follows).

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

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

By (30) and (31), the population slopes and their estimates are

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

Facts:

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

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

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

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

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

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

$$\hat{\beta} = \sum_i w_i\, b_i\,, \qquad b_i := \frac{y_i}{x_i}, \qquad w_i := \frac{x_i^2}{\sum_{i'} x_{i'}^2}, \tag{32}$$

    where b_i and w_i are case-wise slopes and weights, respectively;

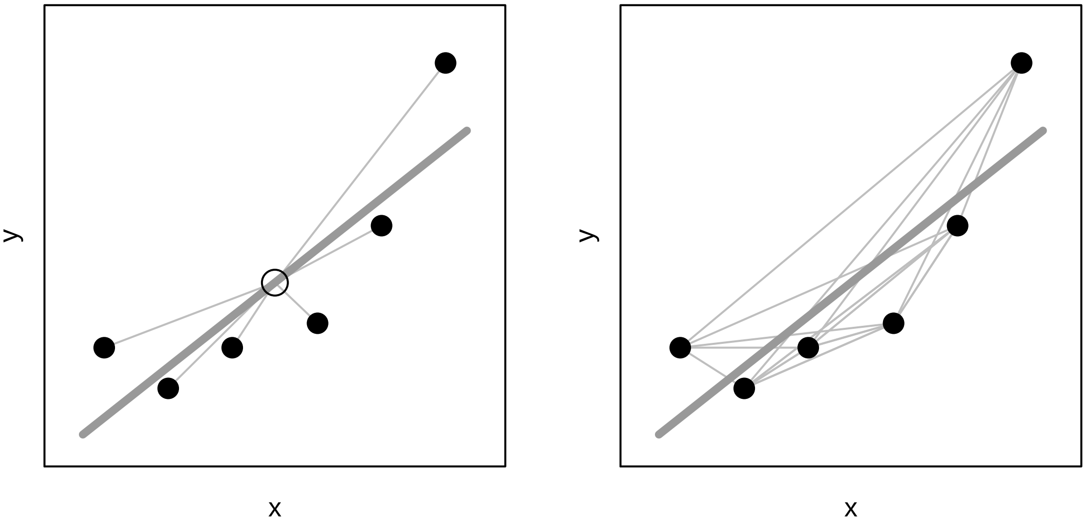

FIG 5. *Case-wise and pairwise average weighted slopes illustrated: Both plots show the same six points ("cases") as well as the OLS line fitted to them (fat gray). The left hand plot shows the case-wise slopes from the mean point (open circle) to the six cases, while the right hand plot shows the pairwise slopes between all 15 pairs. The OLS slope is a weighted average of the case-wise slopes according to* (32)*, and of the pairwise slopes according to* (33)*.*

– **pairwise slopes**:

$$(33)\quad \hat{\beta} = \sum_{ik} w_{ik}\, b_{ik}\,, \qquad b_{ik} := \frac{y_i - y_k}{x_i - x_k}, \qquad w_{ik} := \frac{(x_i - x_k)^2}{\sum_{i'k'} (x_{i'} - x_{k'})^2}\,,$$

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

See Figure 5 for an illustration for samples. The formulas support the intuition that, even in the presence of nonlinearity, a linear fit can be used to infer the overall direction of the association between the response and the regressors.

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

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

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

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

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

This statement is strangely abstract because it refers to an unreal mental scenario of pairs of census tracts that agree in all other regressors but differ in the focal regressor by one unit. Such scenarios are realistic in designed experiments where regressors can be actively manipulated but unrealistic in observational data where regressors are passively received. By comparison the statements (1) and (2) above refer to actual deviations and differences as they are at least potentially observed. Statements (1) and (2) are of course correct in the assumption-laden framework as well. Either way, most end users of models will run with the shorthand "the slope for `PercVacant` is between 4 and 5 homeless per percent."

Note on literature: The above formulas were used and modified to produce alternative slope estimates by Gelman and Park (2008), with the "Goal of Expressing Regressions as Comparisons that can be Understood by the General Reader" (see their Sections 1.2 and 2.2). Earlier, Wu (1986) used generalizations using general tuples rather than pairs of $(\vec{\boldsymbol{x}}_i', y_i)$ rows for the analysis of jackknife and bootstrap procedures (see his Section 3, Theorem 1). The formulas have a history in which Stigler (2001) includes Edgeworth, while Berman (1988) traces it back to a 1841 article by Jacobi written in Latin.

9. ASYMPTOTIC VARIANCES — PROPER AND IMPROPER

The following prepares the ground for an asymptotic comparison of assumption-laden with assumption-lean standard errors. The comparisons will be for one regressor at a time, drawing on the adjustment formalism.

9.1 Preliminaries: Adjustment for Estimation Offsets and Their CLTs

The vectorized formulas for estimation offsets (22) can be written componentwise using adjustment as follows:

$$\begin{array}{lrcl}
\mathit{Total\ EO}: & \hat{\beta}_j - \beta_j & = & \dfrac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\delta} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}, \\[2ex]
\mathit{Noise\ EO}: & \hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] & = & \dfrac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\epsilon} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}, \\[2ex]
\mathit{Nonlinearity\ EO}: & \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j & = & \dfrac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\eta} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}.
\end{array} \tag{34}$$

To see these identities directly, note the following, in addition to (31): $\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}] = \langle \boldsymbol{\mu}, \boldsymbol{X}_{j\hat{\bullet}} \rangle / \|\boldsymbol{X}_{j\hat{\bullet}}\|^2$ and $\beta_j = \langle \boldsymbol{X\beta}, \boldsymbol{X}_{j\hat{\bullet}} \rangle / \|\boldsymbol{X}_{j\hat{\bullet}}\|^2$, the latter due to $\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{X}_k \rangle = \delta_{jk} \|\boldsymbol{X}_{j\hat{\bullet}}\|^2$. Finally use $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X\beta}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X\beta}$ and $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$.

With (34), asymptotic normality of the coefficient-specific EOs can be separately expressed using population adjustment:

Corollary 9.1:

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

The equalities on the right side in the first and second case are based on (24). The first one is needed for plug-in estimation. — Unlike the matrix forms of Proposition 5.4, the univariate asymptotic variances of Corollary 9.1 lend themselves for analyzing individual coefficient estimates. The sandwich form for matrices has been reduced to a ratio where the numerator corresponds to the "meat" and the squared denominator to the "breads".

9.2 Proper Asymptotic Variances in Terms of Adjusted Regressors

The CLTs of Corollary 9.1 contain three asymptotic variances of the same form, the arguments being, respectively, the conditional MSE, the conditional variance, and the squared nonlinearity. This suggests using generic notation:

Definition 9.2:

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

Lemma 9.2: *The proper asymptotic variance of $\hat{\beta}_j$ is*

$$\boldsymbol{AV}^{(j)}_{lean}[m^2(\vec{\boldsymbol{X}})] \;=\; \boldsymbol{AV}^{(j)}_{lean}[\,\sigma^2(\vec{\boldsymbol{X}})] \;+\; \boldsymbol{AV}^{(j)}_{lean}[\eta^2(\vec{\boldsymbol{X}})].$$

9.3 Improper Asymptotic Variances in Terms of Adjusted Regressors

Next we write down an asymptotic form for the standard error estimate from linear models theory in the assumption-lean framework. This asymptotic form will generally be improper in the assumption-lean framework. It derives from an estimate $\hat{\sigma}^2$ of the noise variance, usually $\hat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(N{-}p{-}1)$, which has the following limit for fixed p:

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

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

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \hat{\sigma}^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}, \qquad \hat{\boldsymbol{SE}}^2_{lin}[\,\hat{\beta}_j] \;=\; \frac{\hat{\sigma}^2}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}. \tag{35}$$

Their scaled limits under lean assumptions are as follows:

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

These limits are the "***improper asymptotic variances***" of linear models theory. Again we use a generic definition and an associated decomposition:

Definition 9.3:

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

Lemma 9.3: *The improper asymptotic variance of $\hat{\beta}_j$ in linear models theory is*

$$\boldsymbol{AV}^{(j)}_{lin}[m^2(\vec{\boldsymbol{X}})] \;=\; \boldsymbol{AV}^{(j)}_{lin}[\sigma^2(\vec{\boldsymbol{X}})] \;+\; \boldsymbol{AV}^{(j)}_{lin}[\eta^2(\vec{\boldsymbol{X}})].$$

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

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

Definition 9.4: *Ratio of Asymptotic Variances, Proper/Improper.*

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

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

$$\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] \;=\; w_\sigma\, \boldsymbol{RAV}_j[\sigma^2(\vec{\boldsymbol{X}})] \;+\; w_\eta\, \boldsymbol{RAV}_j[\eta^2(\vec{\boldsymbol{X}})],$$

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

Implications of this decomposition will be discussed below. The three $\boldsymbol{RAV}_j$ terms can be interpreted as inner products between the three random variables

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

These are *not* correlations, and they are not upper bounded by +1; their natural bounds are rather 0 and ∞, both of which can generally be approached to any degree as will be shown in Subsection 9.7.

9.5 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 samples the proper standard error of $\hat{\beta}_j$ is about twice as large as the usual standard error.

If, however, $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] = 1$, it does not follow 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 $\eta(\vec{\boldsymbol{X}}) = 0$. 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 usual standard error.

9.6 Simplification of the *RAV* through Adjustment

The $\boldsymbol{RAV}_j$ is not really a function of all of $\vec{\boldsymbol{X}}$ but of $X_{j\bullet}$ only. Conditioning on $X_{j\bullet}$ reduces $\boldsymbol{RAV}_j$ to a functional of univariate functions:

Lemma 9.6: *Define* $f_j^2(X_{j\bullet}) := \boldsymbol{E}[f^2(\vec{\boldsymbol{X}}) \,|\, X_{j\bullet}{}^2]$. *We have:*

$$\boldsymbol{RAV}_j[f^2(\vec{\boldsymbol{X}})] \;=\; \frac{\boldsymbol{E}[f_j^2(X_{j\bullet})\, X_{j\bullet}{}^2]}{\boldsymbol{E}[f_j^2(X_{j\bullet}]\, \boldsymbol{E}[X_{j\bullet}{}^2]}.$$

For interpretability we think of f_j^2 as a function of $X_{j\bullet}$ rather than $X_{j\bullet}{}^2$, but it should be kept in mind that the actual dependence is on $X_{j\bullet}{}^2$. In general we can write $\boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) = \boldsymbol{RAV}_j(f_j^2(X_{j\bullet}))$, so the analysis of $\boldsymbol{RAV}_j$ can be done in terms of single adjusted regressors $X_{j\bullet}$. This lends itself to simple case studies and graphical illustrations. — Finally we note:

$$m_j^2(X_{j\bullet}) = \eta_j^2(X_{j\bullet}) + \sigma_j^2(X_{j\bullet}).$$

9.7 The Range of *RAV*

We show that for many distributions of $X_{j\bullet}$ the values of $\boldsymbol{RAV}_j$ vary between 0 and ∞. The proposition below states sufficient conditions on the distribution of $X_{j\bullet}$ under which these bounds are sharp. Even though the proposition holds true for $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$ as well, we state it for $m^2(\vec{\boldsymbol{X}})$ which ultimately determines the asymptotic match or mismatch of sandwich and usual standard errors. To examine how widely $\boldsymbol{RAV}_j$ can vary, we consider suprema and infima over scenarios $m^2(\vec{\boldsymbol{X}})$, or rather $m_j^2(X_{j\bullet})$ by Section 9.6.

Proposition 9.7:
(a) If $X_{j\bullet}$ has unbounded support, $\boldsymbol{P}$-max $X_{j\bullet}{}^2 = \infty$, then

$$\sup_{m^2} \boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] \;=\; \infty\,.$$

(b) If $X_{j\bullet}$ has bounded support, $\boldsymbol{P}$-max $X_{j\bullet}{}^2 = c^2 < \infty$, then

$$\sup_{m^2} \boldsymbol{RAV}_j(m^2) = \frac{c^2}{\boldsymbol{E}[X_{j\bullet}{}^2]}.$$

(c) If the mean of $X_{j\bullet}$ $(= 0)$ is in the closure of the support of the distribution of $X_{j\bullet}$ but has zero probability, that is, $\boldsymbol{P}$-min $X_{j\bullet}{}^2 = 0$ but $\boldsymbol{P}[X_{j\bullet}{}^2 = 0] = 0$, then

$$\inf_{m^2} \boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})] = 0\,.$$

According to part (a), the usual standard error can be too small to any degree when the adjusted regressor is unbounded. According to part (b), the usual standard error can be too small by no more than a factor $c^2/\boldsymbol{E}[X_{j\bullet}{}^2]$ if the the adjusted regressor takes on values in the bounded interval $[-c, +c\,]$ only. Part (c) says that the the usual standard error can be too large to any degree when the adjusted regressor takes on values near its mean.

What shapes of $m_j^2(X_{j\bullet})$ approximate these extremes? An intuitive answer can be guessed from Figure 6 for normally distributed $X_{j\bullet}$ to illustrate (a) and (c) of the proposition: If nonlinearities and/or heteroskedasticities blow up ...

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

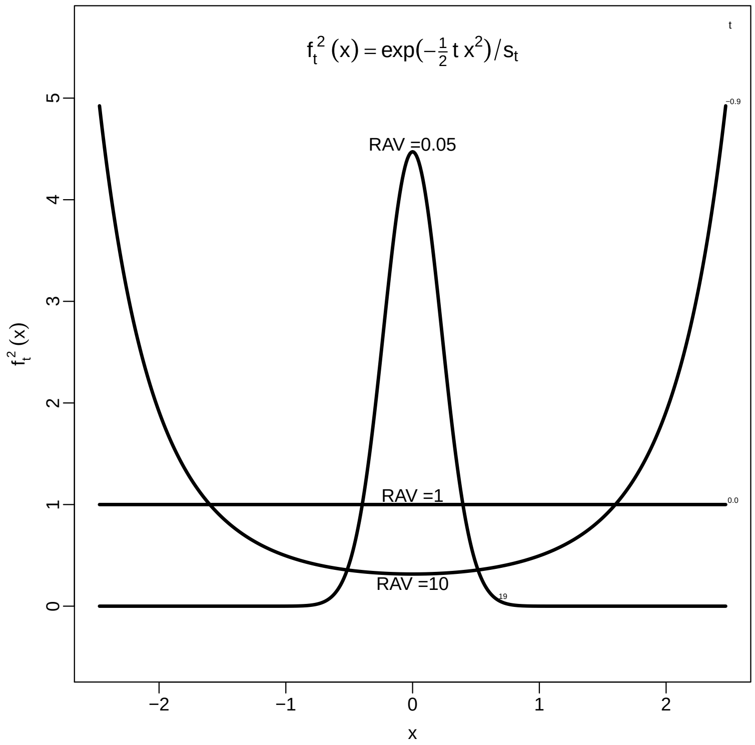

Fig 6. *A family of functions $f_t^2(x)$ that can be interpreted as heteroskedasticities $\sigma_j^2(X_{j\bullet})$, squared nonlinearities $\eta_j^2(X_{j\bullet})$, or conditional MSEs $m_j^2(X_{j\bullet})$: The family interpolates* $\boldsymbol{RAV}$ *from* 0 *to* ∞ *for $x = X_{j\bullet} \sim N(0,1)$.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

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

The proof (in Appendix C.4) bears this out. We are most concerned with the case where the standard errors of linear models theory are too optimistic, that is, most egregiously case (a) of the proposition. Case (b) shows that $X_{j\bullet}$-distributions with bounded support enjoy a degree of protection from the worst case:

- For example, if $X_{j\bullet} \sim U[-1,+1]$ is uniformly distributed, we have $\boldsymbol{E}[{X_{j\bullet}}^2] = 1/3$, hence the upper bound on the $\boldsymbol{RAV}_j$ is 3. It follows that, asymptotically, the usual standard error will never be too short by more than a factor $\sqrt{3} \approx 1.732$.
- However, when $\boldsymbol{E}[{X_{j\bullet}}^2]$ is very small compared to the range of $X_{j\bullet}$, that is, when $X_{j\bullet}$ is highly concentrated around its mean, then this approximates case (a) of the proposition and the worst-case $\boldsymbol{RAV}_j$ can be very large.
- If, on the other hand, $\boldsymbol{E}[{X_{j\bullet}}^2]$ is very close to $\boldsymbol{P}\text{-}max\, {X_{j\bullet}}^2$, it implies that $X_{j\bullet}$ approximates a balanced two-point distribution with probabilities 0.5 at ±1 ($X_{j\bullet}$ must be centered). In this limiting case, the sandwich and usual standard errors necessarily agree in the asymptotic limit.

The result for the last case, a two-point balanced distribution, is intuitive because here it is impossible to detect nonlinearity. On the other hand, heteroskedasticity

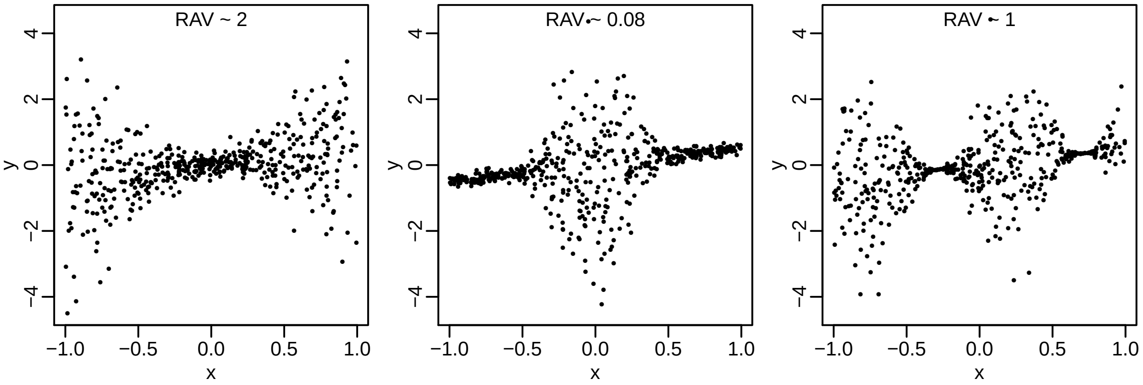

FIG 7. *The effect of heteroskedasticity on the sampling variability of slope estimates: The question is how the misinterpretation of the heteroskedasticities as homoskedastic affects statistical inference.*
Left: High noise variance in the tails of the regressor distribution elevates the true sampling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

is still possible (different noise variances at ±1), but this does not matter because the dependence of $\boldsymbol{RAV}_j$ is on ${X_{j\bullet}}^2$, not $X_{j\bullet}$, and ${X_{j\bullet}}^2$ has a one-point distribution at $+1$. $\boldsymbol{RAV}_j$ can only respond to heteroskedasticities that vary in ${X_{j\bullet}}^2$.

This discussion throws light on the technical condition in part (c): For nonlinearities and heteroskedasticities in the center of the $X_{j\bullet}$ distribution to matter, it is necessary that there is probability mass near that center.

9.8 Illustration of Factors that Drive $\boldsymbol{RAV}_j$

So far we discussed and illustrated the properties of $\boldsymbol{RAV}_j$ in terms of extreme scenarios for $m_j^2(X_{j\bullet})$, which could also be interpreted as scenarios for $\sigma_j^2(X_{j\bullet})$ and $\eta_j^2(X_{j\bullet})$. Next we illustrate in terms of potential data situations: Figure 7 shows three heteroskedasticity scenarios and Figure 8 three nonlinearity scenarios. These examples train our intuitions about the types of heteroskedasticities and nonlinearities that drive the overall $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$. Based on the $\boldsymbol{RAV}_j$ decomposition Lemma 9.4 according to which $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$ is a mixture of $\boldsymbol{RAV}_j[\sigma^2(\vec{\boldsymbol{X}})]$ and $\boldsymbol{RAV}_j[\eta^2(\vec{\boldsymbol{X}})]$, we can state the following:

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

These facts also suggest that in practice, large values $\boldsymbol{RAV}_j > 1$ should occur more often than small values $\boldsymbol{RAV}_j < 1$ because large conditional variances as well as nonlinearities are often more pronounced in the extremes of regressor distributions. This seems particularly natural for nonlinearities which in the simplest

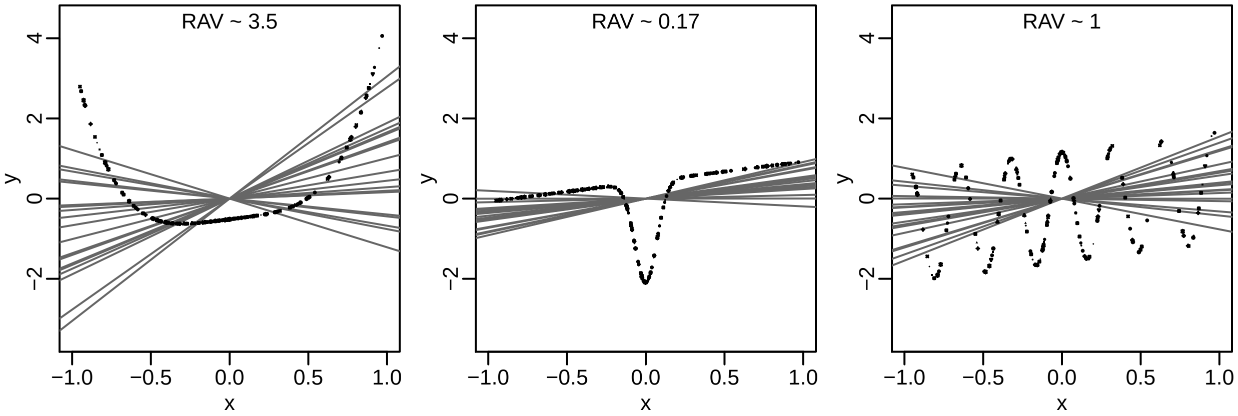

FIG 8. *The effect of nonlinearities on the sampling variability of slope estimates: The three plots show three different noise-free nonlinearities; each plot shows for one nonlinearity 20 overplotted datasets of size $N = 10$ and their fitted lines through the origin. The question is how the misinterpretation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

cases will be convex or concave. In addition it follows from the $\boldsymbol{RAV}$ decomposition Lemma 9.4 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 both have to be close to zero to pull $\boldsymbol{RAV}_j[m^2(\vec{\boldsymbol{X}})]$ toward zero. These considerations are of course mere heuristics for the observation that in practice $\hat{\boldsymbol{SE}}_{lin}$ is more often too small than too large compared to $\hat{\boldsymbol{SE}}_{sand}$.

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

The goal is to write the $\boldsymbol{RAV}$ in adjustment form and estimate it with plug-in for use as a test statistic to decide whether the usual standard error is adequate. In adjustment form we obtain one test per regressor variable.

Two issues will be faced: The asymptotic approximation of the null distribution does not work in practice, and the test is more sensitive to non-normality than to nonlinearity and heteroskedasticity. Both issues will be addressed by a permutation-based approach to the null distribution. Even though the asymptotic null distribution is not practical, it will nevertheless give insight into the detectability of misspecifications that matter for standard error discrepancies.

The proposed test is related to the class of "misspecification tests" for which there exists a literature starting with Hausman (1978) and continuing with White (1980a,b; 1981; 1982) and others. These tests are largely global rather than coefficient-specific, which ours is. The test proposed here has similarities to White's (1982, Section 4) "information matrix test" which compares two types of information matrices globally, while we compare two types of standard errors one coefficient at a time. Another, parameter-specific misspecification test of White (1982, Section 5) compares two types of coefficient estimates rather than standard error estimates, which hence is not a test of standard error discrepancies.

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

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

To begin with, the adjustment versions of the asymptotic variances in the CLTs of Corollary 9.1 can be used to rewrite the sandwich estimator by replacing expectations $\boldsymbol{E}[...]$ with means $\hat{\boldsymbol{E}}[...]$, the population parameter $\boldsymbol{\beta}$ with its estimate $\hat{\boldsymbol{\beta}}$, and population adjustment $X_{j\bullet}$ with sample adjustment $X_{j\hat{\bullet}}$:

$$(36) \qquad \hat{\boldsymbol{AV}}^{(j)}_{sand} \;=\; \frac{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2]}{\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2]^2} \;=\; N\,\frac{\langle(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^4}$$

The squaring of N-vectors is meant to be coordinate-wise. Formula (36) 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 (27) above: $\hat{\boldsymbol{AV}}^{(j)}_{sand} = (\hat{\boldsymbol{AV}}_{sand})_{j,j}$. The sandwich standard error estimate (28) can therefore be written as follows:

$$(37) \qquad \hat{\boldsymbol{SE}}_{sand}(\hat{\beta}_j) \;=\; \frac{\langle(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle^{1/2}}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}.$$

The usual standard error estimate is (35):

$$(38) \qquad \hat{\boldsymbol{SE}}_{lin}(\hat{\beta}_j) \;=\; \frac{\hat{\sigma}}{\|\boldsymbol{X}_{j\hat{\bullet}}\|} \;=\; \frac{\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|}{(N\text{–}p\text{–}1)^{1/2}\,\|\boldsymbol{X}_{j\hat{\bullet}}\|}\,.$$

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 $\boldsymbol{RAV}_j$:

$$(39) \qquad \hat{\boldsymbol{RAV}}_j \;:=\; N\,\frac{\langle(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\;\|\boldsymbol{X}_{j\hat{\bullet}}\|^2} \;=\; \frac{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2\,]}{\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}'\hat{\boldsymbol{\beta}})^2\,]\,\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2\,]}\,.$$

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 $\hat{\boldsymbol{RAV}}_j$ 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 regressors actually matters for judging the effects of misspecification on inference for a particular regression coefficient: it is the squared adjusted regressor and its association with the squared population residuals as estimated by residuals.

10.2 A *RAV* Test for the Discrepancy Between Proper and Improper SEs

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 10.2: *If the population residuals δ_i are independent of $\vec{\boldsymbol{X}}_i$ (not assuming normality of δ_i) we have:*

$$N^{1/2}\,(\boldsymbol{R\hat{A}V}_j - 1) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(0,\, \frac{\boldsymbol{E}[\,\delta^4]}{\boldsymbol{E}[\,\delta^2]^2}\,\frac{\boldsymbol{E}[X_{j\bullet}{}^4]}{\boldsymbol{E}[X_{j\bullet}{}^2]^2} - 1)\right) \tag{40}$$

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 technical assumptions. A proof outline is in Appendix C.6.

According to (40) it is the kurtoses (= the standardized fourth moments - 3) of population residuals δ and of the adjusted regressor $X_{j\bullet}$ that drive the asymptotic variance of $\boldsymbol{R\hat{A}V}_j$ under the null hypothesis. We note the following facts:

1. 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.
2. 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 population residuals δ_i and the adjusted regressor $X_{i,j\bullet}$ have symmetric two-point distributions.
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 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 ways that are 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 population residuals δ and the adjusted regressor $X_{j\bullet}$, a permutation test offers itself. To this end it is necessary that

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

TABLE 4

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

$N \gg p$, and the test will not be exact. The reason is that one needs to estimate the population residuals δ_i with sample residuals r_i and the population adjusted regressor values $X_{i,j\bullet}$ with sample adjusted regressor 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 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 $\boldsymbol{R\hat{A}V}_j$ 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 D for examples).

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

11. ISSUES WITH ASSUMPTION-LEAN STANDARD ERRORS

Model-robustness is a highly desirable property of the sandwich estimator, but as always there is no free lunch:

- As Kauermann and Carroll (2001) have shown, the sandwich estimator may be inefficient when the assumed model is correct. Using plug-in in asymptotic variances can lead to standard errors that are too small/optimistic because the variability from plug-in is not accounted for. Sandwich estimators should therefore be accurate only when the sample size is sufficiently large. This fact suggests that use of the model-dependent standard error should be kept in mind if there is evidence in its favor, for example, through the tests of Section 10.2 and through model diagnostics. (Kauermann and Carroll's analysis is for fixed regressors and treats heteroskedasticity only, but its message is valid because it speaks to performance under well-specification.)
- Another cost associated with the sandwich estimator is non-robustness in the sense of robust statistics (Huber and Ronchetti 2009, Hampel et al. 1986), meaning strong sensitivity to outlying observations: The statistic $\hat{\boldsymbol{SE}}^2_{sand}[\hat{\beta}_j]$ (37) is a ratio of fourth order quantities of the data, whereas $\hat{\boldsymbol{SE}}^2_{lin}[\hat{\beta}_j]$ (38) is "only" a ratio of second order quantities. Note we are

here concerned not with non-robustness of parameter estimates but their standard error estimates. The situation may be no better for the x-y bootstrap standard error because the problem is inherent in the form of the assumption-lean asymptotic variance (Section 9.2) estimated by both sandwich and bootstrap.

According to the second point, two types of robustness are in conflict with each other: Non-robustness to outlying observations arises in standard errors that are sought to be robust to model deviations such as nonlinearities and heteroskedasticities. This is a large issue which we can only raise but not solve in this space. Here are a few observations and suggestions for future research:

- If model-robust standard errors are not classically robust, the converse holds true also: the standard errors of classical robust regression are not model-robust either. In the LA Homeless data for example, for the most important variable `PercVacant`, we observed a ratio of 3.28 comparing the standard error from the x-y bootstrap with the standard error reported by the software (using the function `rlm` in the **R** Language (2008)).
- Yet classical robust regression may confer partial robustness to the sandwich standard error because it limits the size of residuals by capping them with a bounded ψ function. This addresses robustness to outlyingness in the vertical or y direction.
- Robustness to outlyingness in the horizontal or $\vec{\boldsymbol{x}}$ direction could be achieved by using bounded-influence regression (see, e.g., Krasker and Welsch 1982, and references therein) which automatically downweights observations in high-leverage positions, or by using some other downweighting scheme to control the effects of high-leverage points.
- Robustness to horizontal outlyingness could also be addressed by transforming the regressor variables to bounded ranges. Taking a cue from Proposition 9.7, one might search for transformations that obviate the need for an assumption-lean standard error in the first place.

As an illustration of the last point, we transformed the regressors of the LA Homeless data with their empirical cdfs to achieve approximately uniform marginal distributions up to discreteness. The transformed data are no longer i.i.d., but the point is to show the potential effect of transforming the regressors to a finite range and well spread out in it. The results are shown in Table 5: The discrepancies between sandwich and usual standard errors have all but disappeared, and there is no reason to use the former. The same drastic effect is not seen in the Boston Housing data (Appendix A, Table 7), although the discrepancies are greatly reduced here, too. (A technical point is that bounded ranges are really needed for the adjusted regressors, but transformation of the raw regressors is likely to achieve this goal unless the collinearities are extreme.)

12. SUMMARY AND OUTLOOK

Sandwich estimators of standard error are known to be "heteroskedasticity-consistent," meaning that asymptotically they get standard errors right even if there are second order variance deviations from the model. We pointed out that sandwich estimators are also "nonlinearity-consistent", meaning that asymptotically they get standard errors right even if there are first order mean deviations

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

TABLE 5

LA Homeless Data: Comparison of Standard Errors; regressors are transformed with cdfs.

from the model. We analyzed in the simplest case of OLS the joint effect of nonlinearity and heteroskedasticity on sandwich as well as usual standard errors. This effect is mediated by the conditional mean squared error that combines first and second order model deviations. We showed that the usual standard error can fail by arbitrary magnitudes either way, and we described the model departures resulting in such failure: The usual standard error of a slope estimate is

- too small when nonlinearities and/or heteroskedasticities are strong in the *tail* of the adjusted regressor;
- too large when nonlinearities and/or heteroskedasticities are strong in the *center* of the adjusted regressor.

Because nonlinearity is a more severe type of model deviation than heteroskedasticity, we reviewed some fundamentals: When models are approximate, then (1) parameters need to be reinterpreted as statistical functionals on largely arbitrary joint x-y distributions, and (2) the ancillarity principle no longer applies to the regressor distribution. Conclusion (1) requires a reinterpretation of linear slopes in the presence of nonlinearity, and conclusion (2) forces us to treat the regressors as random rather than fixed and to allow that different regressor distributions may result in different slopes. Inference should then rely on assumption-lean/model-robust approaches such as sandwich estimators or the x-y bootstrap, which we showed to be related.

Since White's seminal work, research into misspecification has progressed far and in many forms by addressing specific classes of model deviation: dependencies, heteroskedasticities and nonlinearities. A direct generalization of White's sandwich estimator to time series dependence in regression data is the "heteroskedasticity and auto-correlation consistent" (HAC) estimator of standard error by Newey and West (1987). Structured second order model deviations such as over/underdispersion have been addressed with quasi-likelihood. More generally intra-cluster dependencies in clustered (e.g., longitudinal) data have been addressed with generalized estimating equations (GEE) where the sandwich estimator is in common use, as it is in the generalized method of moments (GMM) literature. Finally, nonlinearities have been modeled with specific function classes or estimated nonparametrically with, for example, additive models, spline and kernel methods, and tree-based fitting.

In spite of these advances it should be kept in mind that in finite data not all possibilities of misspecification can be approached simultaneously, and that therefore a need may arise for assumption-lean inference. It should also be kept in mind that even when complex modeling is possible, simple questions may call for simple models that do not do full justice to the complexity of the data, in which case again it may be wise to look for assumption-lean inference.

There exist, finally, areas of statistics research that appear to be to a large extent assumption-laden:

- Bayes inference, when it relies on uninformative priors, is asymptotically equivalent to assumption-laden frequentist inference. It should not be unreasonable to ask how far inferences from Bayesian models are exposed to adverse effects of model deviation. Complex Bayesian models, however, often use large numbers of fitted parameters and control overfitting by shrinkage, hence asymptotic comparisons may be inadequate and might have to be replaced by finite-sample comparisons in simulations. Interesting developments are taking place, however: Szpiro, Rice and Lumley (2010) derive a sandwich estimator from Bayesian assumptions, and a lively discussion of misspecification from a Bayesian perspective involved Walker (2013), De Blasi (2013), Hoff and Wakefield (2013) and O'Hagan (2013), where further references can be found.
- High-dimensional inference is the subject of a large literature that appears to rely heavily on the assumptions of linearity, homoskedasticity as well as normality of error distributions. It may be uncertain whether procedures proposed in this area are robust to model deviation. Recently, however, attention to misspecification started to be paid by Bühlmann and van de Geer (2015). An interesting development is also the incorporation of ideas from robust statistics by, for example, El Karoui et al. (2013), Donoho and Montanari (2014), and Loh (2015).

Thus there remains some work to be done especially in some of today's most lively research areas to fully realize the consequences from the idea that statistical models are approximations rather than truths.

REFERENCES

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

[2] 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.)

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

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

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

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

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

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

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

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

[11] El Karoui, N. and Bean, D. and Bickel, P. and Yu, B. (2013). Optimal M-estimation in high-dimensional regression. *Proceedings of National Academy of Sciences* **110** (36), 14563-14568.

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

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

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

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

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

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

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

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

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

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

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

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

[24] 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.

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

[26] 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.

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

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

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

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

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

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

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

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

[35] Politis, D. N. and Romano, J. P. (1994). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* **22**, 2031–2050.

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

[37] Sherman, M. and le Cessiea, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models *Communications in Statistics - Simulation and Computation* **26** (3), 901–925.

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

[39] 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.

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

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

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

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

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

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

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

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

TABLE 6
Boston Housing data: Comparison of Standard Errors.

APPENDIX A: THE BOSTON HOUSING DATA

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

Table 7 compares standard errors after the

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

Table 8 illustrates the $\boldsymbol{RAV}$ test for the Boston Housing data. Values of $\boldsymbol{R\hat{A}V}_j$ that fall outside the middle 95% range of their permutation null distributions are marked with asterisks.

APPENDIX B: ANCILLARITY

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

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

TABLE 7

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

	$\hat{\beta}_j$	$\boldsymbol{SE}_{lin}$	$\boldsymbol{SE}_{sand}$	$\hat{\boldsymbol{RAV}}_j$	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 8

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

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

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

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

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

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

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

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

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

APPENDIX C: PROOFS

C.1 Proof of the Lemma in Section 3.3

- 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.
- Population residuals $\delta = \epsilon + \eta$: Again, the conditional distribution must be identical across regressor space, which results in both of the previous cases.

C.2 Proof of the Proposition in Section 4

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

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

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

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

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

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

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

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

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

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

Using $\boldsymbol{E}[\,\delta\vec{\boldsymbol{X}}\,] = \boldsymbol{0}$ from (15) we have:

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

C.4 Proof of the Proposition of Section 9.7

An important difference between $\eta^2(\vec{\boldsymbol{X}})$ and $\sigma^2(\vec{\boldsymbol{X}})$ is that nonlinearities are constrained by orthogonalities to the regressors, 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 regressors: $\boldsymbol{E}[\eta(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}] = \boldsymbol{0}$. To avoid uninsightful complications arising from adjustment due to complex dependencies among the regressors, we construct an example for simple linear regression with a single regressor $X_1 = X$ 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:*

$$\eta_t(X) \;=\; \frac{1_{[|X|>t]} - p(t)}{\sqrt{p(t)(1-p(t))}}\,, \qquad \textit{where} \quad p(t) := \boldsymbol{P}[|X| > t]\,. \tag{41}$$

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:*

$$\sigma_t^2(X) \;=\; \frac{(1_{[|X|>t]} - p(t))^2}{p(t)(1-p(t))}\,, \qquad \textit{where} \quad p(t) = \boldsymbol{P}[|X|>t]\,, \tag{42}$$

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 $1_{[|X|>t]} - p(t) = -1_{[|X|\leq t]} + \bar{p}(t)$:

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

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

C.5 Details for Figure 6

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

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

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

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

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

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

and hence

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

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

C.6 Proof of Asymptotic Normality of $\boldsymbol{R\hat{A}V}_j$, Section 10

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

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

We recall (39) for reference in the following form:

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

For the denominators it is easy to show that

$$\begin{array}{rcl} \frac{1}{N}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 & \xrightarrow{\boldsymbol{P}} & \boldsymbol{E}[\,\delta^2\,], \\ \frac{1}{N}\|\boldsymbol{X}_{j\hat{\bullet}}\|^2 & \xrightarrow{\boldsymbol{P}} & \boldsymbol{E}[\,{X_{j\bullet}}^2\,]. \end{array} \tag{44}$$

For the numerator a CLT holds based on

$$(45)\quad \tfrac{1}{N^{1/2}}\langle(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle \;=\; \tfrac{1}{N^{1/2}}\langle(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^2, \boldsymbol{X}_{j\bullet}{}^2\rangle + O_P(N^{-1/2}).$$

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:

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

The second equality on each line holds under the null hypothesis of independent δ and $\vec{\boldsymbol{X}}$. For the variance one observes that we assume that $\{(Y_i, \vec{\boldsymbol{X}}_i)\}_{i=1...N}$ to be i.i.d. sampled pairs, hence $\{(\delta_i^2, X_{i,j\bullet}{}^2)\}_{i=1...N}$ are N i.i.d. sampled pairs as well. Using the denominator terms (44) and Slutsky's theorem, we arrive at the first version of the CLT for $\boldsymbol{R\hat{A}V}_j$:

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

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 $\boldsymbol{R\hat{A}V}_j$:

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

Details for the numerator (45), 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 $\boldsymbol{X}_{j\hat{\bullet}} = \boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}}$:

$$(46)$$

$$\begin{array}{rcl} \langle(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle &=& \langle\,((\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2,\, (\boldsymbol{X}_{j\bullet} - \boldsymbol{X}_{-j}(\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} - \boldsymbol{\beta}_{-j\bullet}))^2\,\rangle \\ &=& \langle\,\boldsymbol{\delta}^2 + (\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2 - 2\,\boldsymbol{\delta}\,(\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})), \\ && \quad \boldsymbol{X}_{j\bullet}{}^2 + (\boldsymbol{X}_{-j}(\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} - \hat{\boldsymbol{\beta}}_{-j\bullet}))^2 - 2\,\boldsymbol{X}_{j\bullet}(\boldsymbol{X}_{-j}(\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}} - \boldsymbol{\beta}_{-j\bullet}))\,\rangle \\ &=& \langle\,\boldsymbol{\delta}^2, \boldsymbol{X}_{j\bullet}{}^2\,\rangle + ... \end{array}$$

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 $N^{1/2}$:

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

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

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

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

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

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

We proceed to the 6 terms in (46) 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 (44): It is sufficient to consider the first denominator term. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ be the hat or projection matrix for $\boldsymbol{X}$.

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

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

APPENDIX D: NON-NORMALITY OF CONDITIONAL NULL DISTRIBUTIONS OF $\widehat{RAV}_J$

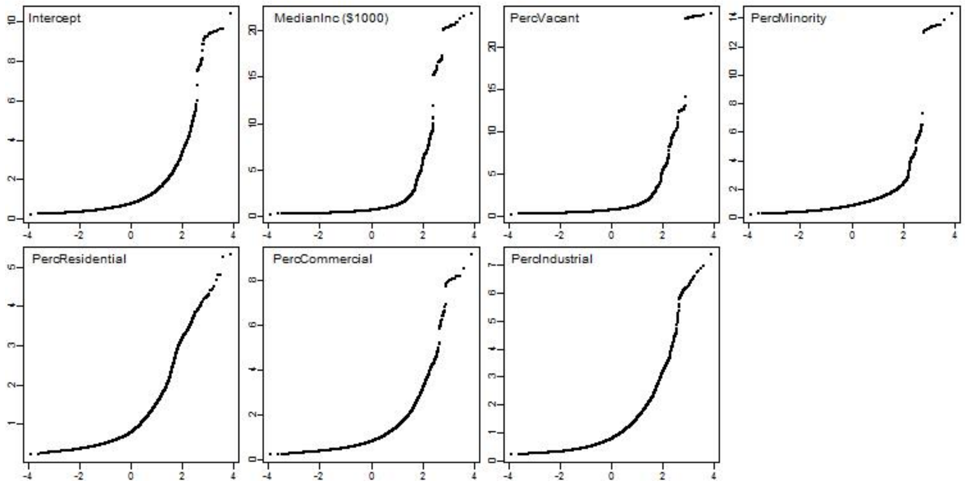

FIG 9. *Permutations distributions of* $\widehat{\boldsymbol{RAV}}_j$ *for the LA Homeless Data*

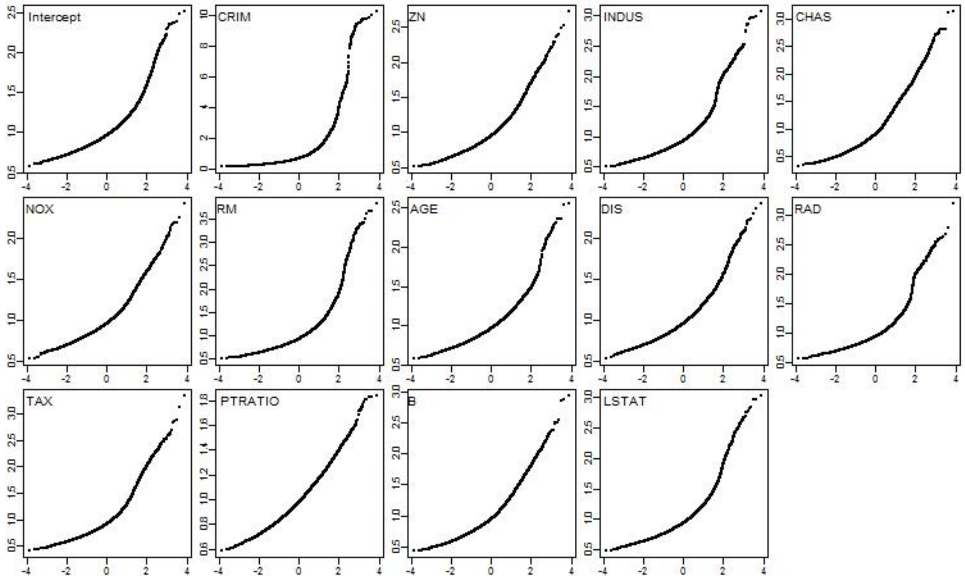

FIG 10. *Permutations distributions of* $\widehat{\boldsymbol{RAV}}_j$ *for the Boston Housing Data*