

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

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

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

Dedicated to Halbert White (†2012)

Abstract.

More than thirty years ago Halbert White inaugurated a “model-robust” form of statistical inference based on the “sandwich estimator” of standard error. This estimator is known to be “heteroskedasticity-consistent”, but it is less well-known to be “nonlinearity-consistent” as well. Nonlinearity raises fundamental issues because regressors are no longer ancillary, hence can’t be treated as fixed. As a result, (1) the regressor distribution affects the parameters and (2) randomness of the regressors conspires with the nonlinearity to become a source of sampling variability in coefficient estimates. These effects generalize to arbitrary types of regression where regressors have traditionally been treated as ancillary. The generalizations result in a novel notion of misspecification and a re-interpretation of regression parameters as statistical functionals. The cost of a model-robust approach is that the meaning of parameters needs to be rethought and inference needs to be based on model-robust standard errors. For linear OLS, model-trusting standard errors can deviate from “model-robust” standard errors by arbitrary magnitudes. In practice, the two types of standard errors can be compared with a diagnostic test.

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

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

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

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

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

1. INTRODUCTION

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

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

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

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

where $\hat{\mathbf{D}}$ is diagonal with $\hat{\mathbf{D}}_{ii} = r_i^2$. Standard error estimates are obtained by $\hat{\mathbf{S}}\mathbf{E}_{sand}[\hat{\beta}_j] = \hat{\mathbf{V}}_{sand}[\hat{\beta}]_{jj}^{1/2}$. 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:

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

The term “heteroskedasticity-consistent” is an unfortunate choice as it obscures the fact that the same estimator of standard error is also “nonlinearity-consistent” when the regressors are random. Because of the relative obscurity of this important fact we will pay considerable attention to its implications. In particular we show how nonlinearity “conspires” with randomness of the regressors (1) to make slopes dependent on the regressor distribution and (2) to generate sampling variability all of its own even in the absence of noise; see Figures 2 and 4 below. A more striking illustration is available to users of the **R** *Language* by executing the following line of code:

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

Side remarks:

- The term “nonlinearity” is meant in the sense of (3), first order model misspecification. A different meaning of “nonlinearity”, *not* intended here, occurs when the regressor matrix \mathbf{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). Some forms are pervasive in Generalized Estimating Equations (GEE; Liang and Zeger 1986; Diggle et al. 2002) and in the Generalized Method of Moments (GMM; Hansen 1982; Hall 2005).

when the linear model is first order correct

when the regressor vectors are iid random variables

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

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

This yields finite-sample unbiased squared standard error estimators $\hat{SE}_{lin}^2[\hat{\beta}_j] = \hat{V}_{in}[\hat{\beta}]_{jj}$ if the model is first and second order correct: $\mathbf{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$ (linearity) and $\mathbf{V}[\mathbf{y}|\mathbf{X}] = \sigma^2\mathbf{I}_N$ (homoskedasticity). Assuming distributional correctness (Gaussian errors), 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 (y_i, \mathbf{x}'_i) of the data matrix (\mathbf{y}, \mathbf{X}) are i.i.d. samples from a joint multivariate distribution with finite moments to some order. Thus sandwich-based theory provides asymptotically correct inference that is **model-robust**. The question then arises what model-robust inference is about: When no model is assumed, *what are the parameters*, and *what is their meaning*?

Answering these and related questions is a first goal of the present article. An established answer is that parameters can be interpreted as *statistical functionals* $\beta(\mathbf{P})$ defined on a large nonparametric class of joint distributions $\mathbf{P} = \mathbf{P}(dy, d\mathbf{x})$ through best approximation (Section 3). The sandwich estimator produces then asymptotically correct standard errors for the slope functionals $\beta_j(\mathbf{P})$ (Section 7). The question of the meaning of slopes in the presence of nonlinearity will be answered with proposals involving *case-wise and pairwise slopes* (Section 9.2).

A second goal of this article is to discuss the role of the regressors when they are random. Based on an ancillarity argument, model-trusting theory tends to condition on the regressors and treat them as fixed. It will be shown, however, that in a model-robust theory the ancillarity principle is generally violated: population parameters depend on the distribution of the regressors (Section 5). In fact, we will propose a generalized notion of well-specification for statistical functionals based on the condition that regressor distributions do not affect them (Section 6).

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

A fourth goal of this article is to practically (Section 2) and theoretically (Section 9.3) compare model-robust and model-trusting estimators. We define a ratio of asymptotic variances — “**RAV**” for short — that describes the discrepancies between the two standard errors in the asymptotic limit. If **RAV** $\neq 1$, it is model-robust estimators (sandwich or *x-y* bootstrap) that are asymptotically correct, and the usual model-trusting standard error is indeed asymptotically incorrect. The **RAV** can range from 0 to ∞ under scenarios that illustrate how model deviations can invalidate the usual standard error.

A fifth goal is to estimate the **RAV** for use as a test statistic. We derive an asymptotic null distribution to test for model deviations that invalidate the usual standard error of a specific coefficient. The resulting “misspecification test” differs from other such tests in that it answers the question of discrepancies

among standard errors directly and separately for each coefficient (Section 9.4). It should be noted that there are “misspecifications” that do not invalidate the usual model-trusting standard error.

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

The level of generality of treatment will vary. Greatest generality, at the level of arbitrary statistical functionals, will apply to foundational topics such as targets of estimation, regressor non-ancillarity, the meaning of mis/well-specification, but also the connection between plug-in/sandwich estimators and x - y bootstrap. Other topics, such as comparisons of model-trusting and model-robust standard errors, will be developed only for linear OLS. The reason is that for these topics the most lucid presentation relies on OLS regressor adjustment, which permits reducing the analysis to 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. Readers may browse the tables and figures and read associated sections that seem most germane. Important notations are shown in boxes.

The idea that models are approximations and hence generally “misspecified” to a degree has a long history, most famously expressed by Box (1979). We prefer to quote Cox (1995): “it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization.” The history of inference under misspecification can be traced to Cox (1961, 1962), Eicker (1963), Berk (1966, 1970), Huber (1967), before being systematically elaborated by White in a series of articles (White 1980a, 1980b, 1981, 1982, among others) and capped by a monograph (White 1994). More recently, a wide-ranging discussion by Wasserman (2011) calls for “Low Assumptions, High Dimensions.” A book by Davies (2014) elaborates the idea of adequate models for a given sample size. We, the present authors, got involved with this topic through our work on post-selection inference (Berk et al. 2013) because the results of model selection should certainly not be assumed to be “correct.” We compared the obviously model-robust standard errors of the x - y bootstrap with the usual ones of linear models theory and found the discrepancies illustrated in Section 2. Attempting to account for these discrepancies became the starting point of the present article.

2. DISCREPANCIES BETWEEN STANDARD ERRORS ILLUSTRATED

Table 1 shows regression results for a dataset consisting of a sample of 505 census tracts in Los Angeles that has been used to examine homelessness in relation to covariates for demographics and building usage (Berk et al. 2008). We do not intend a careful modeling exercise but show the raw results of linear regression to illustrate the degree to which discrepancies can arise among three types of standard errors: \mathbf{SE}_{lin} from linear models theory, \mathbf{SE}_{boot} from the x - y bootstrap ($N_{boot} = 100,000$) and \mathbf{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 $\mathbf{SE}_{sand}/\mathbf{SE}_{boot}$ show that the sandwich and bootstrap estimators are in good agreement. Not so for the linear models estimates: we have $\mathbf{SE}_{boot}, \mathbf{SE}_{sand} >$

	$\hat{\beta}_j$	SE_{lin}	SE_{boot}	SE_{sand}	$\frac{SE_{boot}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{boot}}$	t_{lin}	t_{boot}	t_{sand}
Intercept	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.

SE_{lin} for the regressors PercVacant, PercCommercial and PercIndustrial, and $SE_{boot}, SE_{sand} < SE_{lin}$ for Intercept, MedianInc (\$1000), PercResidential. Only for PercMinority is SE_{lin} off by less than 10% from SE_{boot} and SE_{sand} . The discrepancies affect outcomes of some of the t -tests: Under linear models theory the regressors PercCommercial and PercIndustrial have commanding t -values of 2.700 and 2.818, respectively, which are reduced to unconvincing values below 1.9 and 1.6, respectively, if the x - y bootstrap or the sandwich estimator are used. On the other hand, for MedianInc (\$K) the t -value -0.977 from linear models theory becomes borderline significant with the bootstrap or sandwich estimator if the plausible one-sided alternative with negative sign is used.

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

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

3. THE POPULATION FRAMEWORK

3.1 Populations for Regression

As mentioned in the introduction, parameters of generative models will be reinterpreted as statistical functionals that are well-defined for a large nonparametric class of data distributions. In an assumption-lean, model-robust population framework for regression with random regressors, the ingredients are regressor random variables X_1, \dots, X_p and a response random variable Y . For now the only assumption is that they have a joint distribution, written as

$$P = P(dy, dx_1, \dots, dx_p).$$

For general considerations these random variables need not be quantitative; they may be categorical, ordinal, censored, vector, For specific models they will be limited to grant identifiability of the parameters, as when linear models of any kind require quantitative regressors with second moments and a full-rank regressor covariance. In their case it is convenient to prepend a fixed regressor 1 to accommodate an intercept parameter. We may then write

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

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

$$P = P(dy, d\vec{x}), \quad P(dy | \vec{x}), \quad P(d\vec{x}), \quad \text{or} \quad P = P_{Y, \vec{X}}, \quad P_{Y | \vec{X}}, \quad P_{\vec{X}},$$

for, respectively, the joint distribution of (Y, \vec{X}) , the conditional distribution of Y given \vec{X} , and the marginal distribution of \vec{X} .

In linear models with an intercept, the regressor distribution $P_{\vec{X}}$ is trivially degenerate in \mathbb{R}^{p+1} . There may arise nonlinear degeneracies if multiple regressors are functions of one underlying independent variable, as in polynomial or B-spline regression or product interactions. These cases of degeneracies are permitted as long as $E[\vec{X}\vec{X}']$ remains full-rank.

3.2 Targets of Estimation 1: The Linear OLS Statistical Functional

For linear OLS we assume Y and all X_j quantitative. We write any function $f(X_1, \dots, X_p)$ of the regressors as $f(\vec{X})$ because a prepended constant 1 is irrelevant. The following functions of \vec{X} are special:

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

$$(4) \quad \mu(\vec{X}) := \operatorname{argmin}_{f(\vec{X}) \in L_2(\mathbf{P})} E[(Y - f(\vec{X}))^2] = E[Y | \vec{X}].$$

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

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

$$(5) \quad \beta(\mathbf{P}) := \operatorname{argmin}_{\beta \in \mathbb{R}^{p+1}} E[(Y - \beta' \vec{X})^2] = E[\vec{X}\vec{X}']^{-1} E[\vec{X}Y]$$

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

The right hand expressions follow from the population normal equations:

$$(7) \quad E[\vec{X}\vec{X}']\beta - E[\vec{X}Y] = E[\vec{X}\vec{X}']\beta - E[\vec{X}\mu(\vec{X})] = \mathbf{0}.$$

The population coefficients $\beta = \beta(\mathbf{P})$ form a *vector statistical functional* defined for a large class of joint data distributions $\mathbf{P} = P_{Y, \vec{X}}$.

3.3 Targets of Estimation 2: ML and MoM Statistical Functionals

A model-robust interpretation in terms of statistical functionals can be given to large classes of regression methods for arbitrary variable types:

- **Maximum likelihood (ML):** Given a regression model $p(y | \vec{x}; \theta)$ define a statistical functional by minimization,

$$(8) \quad \theta(\mathbf{P}) = \operatorname{argmin}_{\theta} E_{\mathbf{P}}[-\log p(Y | \vec{X}; \theta)],$$

or by solving the associated moment conditions/estimating equations,

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

Under mild regularity conditions we have $\theta(\mathbf{P}) = \theta_0$ if the actual conditional data distribution $P_{Y|\vec{X}}$ has density $p(y | \vec{x}; \theta_0)$. The point is, however, that $\theta(\mathbf{P})$ is defined for a large class of data distributions outside of the model $p(y | \vec{x}; \theta)$. Models have here a two-fold role:

- To provide a heuristic for an objective function: $\mathcal{L}(\theta; y, \vec{x}) = -\log p(y | \vec{x}; \theta)$.

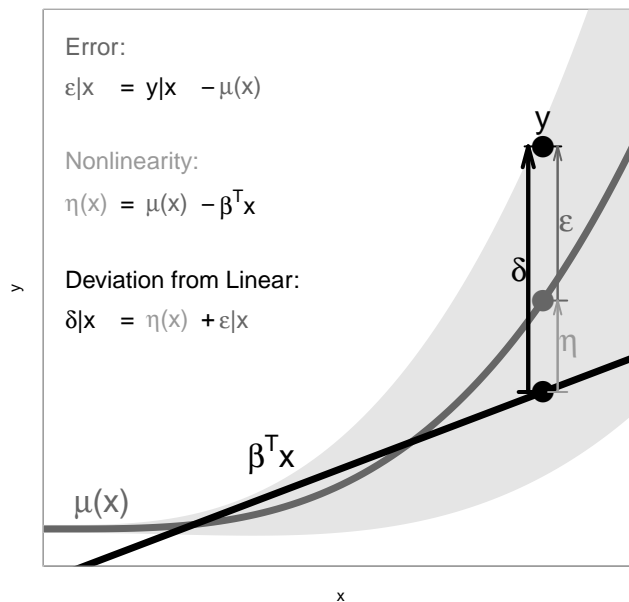


FIG 1. Illustration of the decomposition (13) for linear OLS.

- For the model $p(y | \vec{x}; \theta)$ to act as an approximation to the actual conditional data distribution $P_{Y|\vec{X}}$ (an early adopter being Kent (1982)).
- Generalizing further one may define statistical functionals from objective functions $\mathcal{L}(\theta; y, \vec{x})$ that are not necessarily negative log-likelihoods of a model:

$$(10) \quad \theta(\mathbf{P}) = \operatorname{argmin}_{\theta} \mathbf{E}_{\mathbf{P}}[\mathcal{L}(\theta; Y, \vec{X})].$$

An example is conditional quantile estimation based on tilted L_1 losses.

- *Method of Moments* (MoM): The minimization problem (10) is usually solved in terms of stationarity conditions that amount to moment conditions for $\psi(\theta; y, \vec{x}) = \partial_{\theta} \mathcal{L}(\theta; Y, \vec{X})$:

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

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

$$(12) \quad \psi_{OLS}(\beta; y, \vec{x}) = \vec{x}\vec{x}'\beta - \vec{x}y.$$

- An extension to situations where the number of moment conditions (the dimension of ψ) is larger than the dimension of θ is provided by the Generalized Method of Moments (GMM, Hansen 1982) which can be used for causal inference based on numerous instrumental variables.

Moment conditions for clustered data with intra-cluster dependence are provided by Generalized Estimating Equations (GEE, Liang and Zeger 1986). This, however, is a “fixed- \mathbf{X} ” approach that assumes well-specification of the mean function while allowing misspecification of variance and intra-cluster dependence.

Just the same, it is evident that model-robust interpretations exist for many regression methods. The point of view is to interpret regression parameters as statistical functionals. Accordingly, some of the following discussions will involve general statistical functionals $\theta = \theta(\mathbf{P}_{Y, \vec{\mathbf{X}}})$ in the context of the special structure afforded by the distinction between response and regressor variables.

4. THE NOISE-NONLINEARITY DECOMPOSITION FOR LINEAR OLS

First, we briefly treat linear OLS for its explicit formulas used throughout for illustration. The response Y has the following decompositions:

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

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. Important to note is that (13) is a decomposition; it makes no model assumptions on δ or ϵ . In a model-robust framework with random regressors there is no notion of “error term” in the usual sense; its place is taken by the population residual δ which satisfies few of the usual assumptions made in generative models. It naturally decomposes into a systematic component, the nonlinearity $\eta(\vec{\mathbf{X}})$, and a random component, the noise ϵ . In model-trusting linear modeling, one assumes $\eta(\vec{\mathbf{X}}) \stackrel{P}{=} 0$ and ϵ to have the same $\vec{\mathbf{X}}$ -conditional distribution in all of predictor space, that is, ϵ is assumed independent of $\vec{\mathbf{X}}$ if the latter is treated as random. No such assumptions are made here. What is left are orthogonality conditions satisfied by η and ϵ in relation to $\vec{\mathbf{X}}$. If we call independence “strong-sense orthogonality”, we have instead

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

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

In what follows we will need some natural definitions:

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

$$(15) \quad \sigma^2(\vec{\mathbf{X}}) := \mathbf{V}[\epsilon | \vec{\mathbf{X}}] = \mathbf{E}[\epsilon^2 | \vec{\mathbf{X}}] \stackrel{P}{<} \infty.$$

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

TABLE 2
Random variables and their canonical decompositions.

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

$$(16) \quad m^2(\vec{X}) := E[\delta^2 | \vec{X}] = \eta^2(\vec{X}) + \sigma^2(\vec{X}).$$

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

- **Marginal Noise Variance:** Averaging $m^2(\vec{X})$, $\eta^2(\vec{X})$ and $\sigma^2(\vec{X})$ we could define three second order functionals of \mathbf{P} , but among them we will only need the marginal noise variance:

$$(17) \quad \sigma^2(\mathbf{P}) := E[\sigma^2(\vec{X})] = E[\epsilon^2].$$

5. NON-ANCILLARITY OF THE REGRESSOR DISTRIBUTION

In this section we use linear models for illustration, but the effects are general. Section 6 will describe generalizations to arbitrary statistical functionals.

5.1 The Breakdown of the Ancillarity Argument under Misspecification

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:

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

referring to model densities of $P_{\vec{X}, Y}$, $P_{Y | \vec{X}}$ and $P_{\vec{X}}$, respectively, where θ is the parameter in the traditional meaning of a parametric model. While θ is the parameter of interest, the regressor density $p(\vec{x})$ acts as a “nonparametric nuisance parameter.” Ancillarity of $p(\vec{x})$ in relation to θ is immediately recognized by forming likelihood ratios $p(y, \vec{x}; \theta_1) / p(y, \vec{x}; \theta_2) = p(y | \vec{x}; \theta_1) / p(y | \vec{x}; \theta_2)$ which are free of $p(\vec{x})$. (For a fuller definition of ancillarity see Appendix B.) This logic is valid if the conditional model $p(y | \vec{x}; \theta)$ is correct. The following proposition describes for linear models the ways in which ancillarity is broken if the model is an approximation and the parameters are statistical functionals:

Proposition 5.1: Breaking Regressor Ancillarity in linear OLS

- *Considering distributions $\mathbf{P} = P_{Y, \vec{X}}$ that share the function $\mu(\vec{x})$ as conditional expectation of the response, the functional $\beta(\mathbf{P})$ depends on the regressor distribution $P_{\vec{X}}$ if and only if $\mu(\vec{x})$ is nonlinear.*

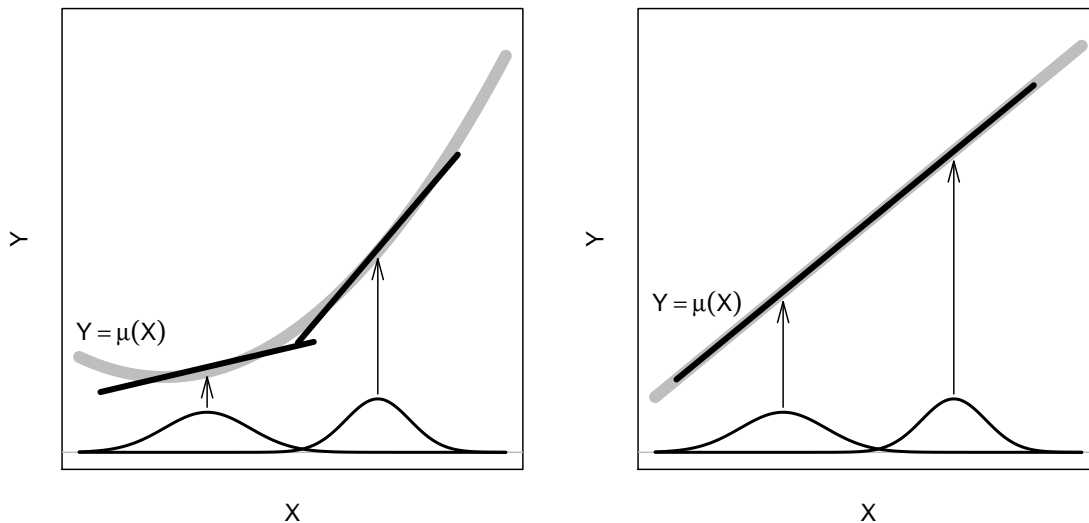


FIG 2. Illustration of the dependence of the population OLS solution on the marginal distribution of the regressors: The left figure shows dependence in the presence of nonlinearity; the right figure shows independence in the presence of linearity.

- Considering distributions $\mathbf{P} = \mathbf{P}_{Y, \bar{\mathbf{x}}}$ that share the function $\sigma^2(\bar{\mathbf{x}})$ as conditional variance of the response, the functional $\sigma^2(\mathbf{P})$ depends on the regressor distribution $\mathbf{P}_{\bar{\mathbf{x}}}$ if and only if $\sigma^2(\bar{\mathbf{x}})$ is non-constant (heteroskedastic).

(These are loose statements; see Appendix D.1 for more precision.) The first part of the proposition is best explained graphically: 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. (See also White (1980a, p. 155f); identify his $g(Z) + \epsilon$ with our Y .)

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

5.2 Implications of the Dependence of Slopes on Regressor Distributions

A first practical implication, illustrated by Figure 2, is that two empirical studies that use the same regressors, the same response, and the same model, may yet estimate different parameter values, $\beta(\mathbf{P}_1) \neq \beta(\mathbf{P}_2)$. What may seem to be superficially contradictory inferences from the two studies may be compatible if (1) the true response surface $\mu(\bar{\mathbf{x}})$ 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 $p \rightarrow \infty$. Differences in estimated parameter values often become visible in meta-analyses and are labeled “parameter heterogeneity.” The source of this heterogeneity may be differences in regressor distributions combined with model misspecification.

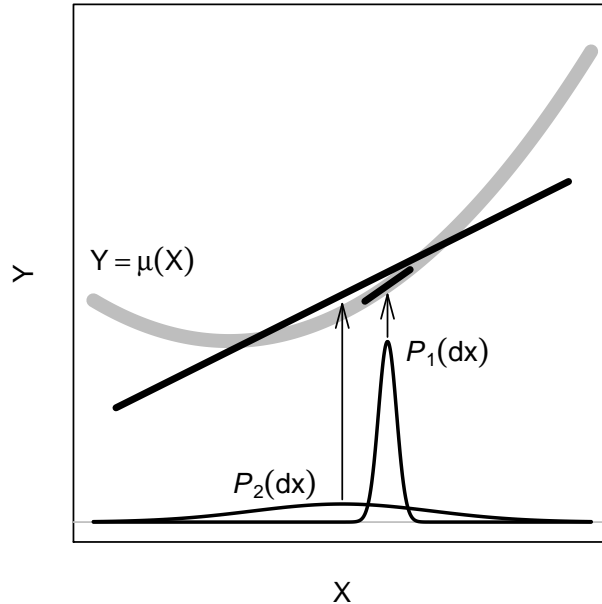


FIG 3. *Illustration of the interplay between regressors' high-density range and nonlinearity: Over the small range of P_1 the nonlinearity will be undetectable and immaterial for realistic sample sizes, whereas over the extended range of P_2 the nonlinearity is more likely to be detectable and relevant.*

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 $P_1(d\vec{x})$ is the reason why the linear approximation is excellent, hence the model very nearly well-specified, whereas the wide range of $P_2(d\vec{x})$ 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.

delete these two sentences

6. REGRESSOR ANCILLARITY AS THE BASIS FOR A NOTION OF WELL-SPECIFICATION OF STATISTICAL FUNCTIONALS

The discussions of the previous section suggest generalizations to arbitrary functionals in regressions with arbitrary data types. We will take Proposition 5 as a guide: from the “only if” parts of its statements we infer that well-specification can be characterized by the condition that statistical functionals be independent of the regressor distribution. For slope functionals, well-specification in this sense implies that the conditional expectation of the response is a linear function of the regressors (linearity), and for the noise variance functional it implies that the conditional variance of the response is constant across the regressor range (homoskedasticity). The condition that a functional be independent of the regressor distribution, however, is so completely non-specific that it lends itself as a definition of well-specification for arbitrary statistical functionals in arbitrary types of regressions.

Beyond Proposition 5 this definition can be further supported by the following general heuristic argument: A statistical functional, in the context of regression, ideally measures aspects of the association between Y and \vec{X} only, not aspects of the marginal distribution of \vec{X} . Therefore, ideally, a statistical functional should not depend on the marginal \vec{X} distribution.

It is important to view such a definition as stating an ideal, not the reality in most data analyses, but ideals are useful because they spell out the circumstances under which intentions are perfectly realized. Here the intention is that measurements of aspects of the Y - \vec{X} association be independent of where the measurements are taken, that is, independent of the \vec{X} distribution. This intention is not realized for slope functionals when the true conditional response mean $\mu(\vec{x})$ is nonlinear; in this case a linear fit will result in larger slopes when the \vec{X} distribution concentrates in locations where the gradient of $\mu(\vec{x})$ is steeper.

The definition proposed here can be interpreted as shifting the concept of regressor ancillarity from models to statistical functionals and using it as the basis for a novel concept of well-specification. This makes sense in light of a comment made by a reviewer of this article who characterized ancillarity as “an example of how much of classic parametric statistics (including concepts such as ancillarity, sufficiency, invariance, exact pivotal quantities) developed up to about 1960 mostly assumes correctness of the working model. This is of less (some might say no) relevance when the correctness of a model is not assumed.” Yet, there is value in leveraging an “outmoded” concept to characterize the ideal circumstances for measuring aspects of a Y - \vec{X} association. For one thing, well-specification of statistical functionals holds for much larger classes of data distributions than those contained in traditionally assumed models. For example, slope functionals are well-specified iff the conditional response mean is a linear function of the regressors — a property which characterizes a large nonparametric class of distributions much beyond linear models with their additional assumptions such as homoskedasticity and Gaussianity.

In what follows we will give precision to the notion of well-specification for statistical functionals. If $\theta(\mathbf{P})$ is a statistical functional defined for a joint distribution $\mathbf{P} = \mathbf{P}_{Y, \vec{X}}$, we may decompose $\mathbf{P}_{Y, \vec{X}}$ into two components: the conditional response distribution $\mathbf{P}_{Y|\vec{X}}$ and the marginal regressor distribution $\mathbf{P}_{\vec{X}}$. In general, conditional distributions are defined only almost surely w.r.t. $\mathbf{P}_{\vec{X}}$, but for technical reasons we will assume that $\vec{x} \mapsto \mathbf{P}_{Y|\vec{X}=\vec{x}}$ is a Markov kernel defined for all $\vec{x} \in \mathcal{X}$, where $\mathcal{X} = \text{supp}(\mathbf{P}_{\vec{X}})$ is the (topologically closed) support of the regressor distribution. We can then write

$$(18) \quad \theta(\mathbf{P}_{Y, \vec{X}}) = \theta(\mathbf{P}_{Y|\vec{X}}, \mathbf{P}_{\vec{X}}).$$

Definition: *The statistical functional $\theta(\mathbf{P})$ is well-specified for $\mathbf{P} = \mathbf{P}_{Y, \vec{X}}$ if*

$$\theta(\mathbf{P}_{Y|\vec{X}}, \mathbf{P}_{\vec{X}}) = \theta(\mathbf{P}_{Y|\vec{X}}, \tilde{\mathbf{P}}_{\vec{X}})$$

for all permissible regressor distributions $\tilde{\mathbf{P}}_{\vec{X}}$ with $\text{supp}(\tilde{\mathbf{P}}_{\vec{X}}) \subset \text{supp}(\mathbf{P}_{\vec{X}})$.

We added the term “permissible” to account for exclusions such as perfectly collinear regressor distributions in linear fitting.

We turn next to special cases and general implications:

- The definition divorces the notion of well-specification from statistical models and makes it a relational property between conditional response distributions, $\mathbf{P}_{Y|\vec{\mathbf{X}}}$, and quantities of interest, $\boldsymbol{\theta}$. Parametric models are replaced by large nonparametric sets of conditional distributions associated with a statistical functional $\boldsymbol{\theta}(\mathbf{P})$:

$$\mathcal{P}_{\boldsymbol{\theta}} := \{ \mathbf{P}_{Y|\vec{\mathbf{X}}} \mid \boldsymbol{\theta} \text{ is well-defined for } \mathbf{P}_{Y,\vec{\mathbf{X}}} \}.$$

(Again we assume $\mathbf{P}_{Y|\vec{\mathbf{X}}=\vec{\mathbf{x}}}$ to be actual Markov kernels defined for all $\vec{\mathbf{x}}$.) For the slope functional $\boldsymbol{\beta}(\mathbf{P})$ of linear OLS, this would be the set of conditional distributions (Markov kernels) for which the conditional expectation is linear in $\vec{\mathbf{X}}$: $\mathbf{E}[Y|\vec{\mathbf{X}}] = \boldsymbol{\beta}(\mathbf{P})'\vec{\mathbf{X}}$.


- If the statistical functional $\boldsymbol{\theta}$ is the population version of an ML estimator for a particular regression model as described in Section 3.3, the functional will be well-specified for the distributions in this model in the sense of the above definition: if $\mathbf{P}_{Y|\vec{\mathbf{X}}=\vec{\mathbf{x}}}$ has conditional density $p(y|\vec{\mathbf{x}};\boldsymbol{\theta}_0)$, then

$$\mathbf{E}_{\mathbf{P}}[-\log p(Y|\vec{\mathbf{X}};\boldsymbol{\theta}_0) \mid \vec{\mathbf{X}}=\vec{\mathbf{x}}] = \min_{\boldsymbol{\theta}} \mathbf{E}_{\mathbf{P}}[-\log p(Y|\vec{\mathbf{X}};\boldsymbol{\theta}) \mid \vec{\mathbf{X}}=\vec{\mathbf{x}}].$$

This holds conditionally on $\vec{\mathbf{X}}=\vec{\mathbf{x}}$, hence it holds marginally irrespective of $\mathbf{P}_{\vec{\mathbf{X}}}$. (Minimizing the r.h.s. is generally highly non-unique; e.g., in linear OLS there exist many linear functions through $(\vec{\mathbf{x}}, \mathbf{E}[Y|\vec{\mathbf{X}}=\vec{\mathbf{x}}])$.)

- Similar arguments apply to non-ML functionals such as quantile regressions. A linear quantile regression, for example, is well-specified if the conditional response quantiles follow a linear function of $\vec{\mathbf{X}}$, meaning that the values of the linear function minimize the associated tilted L_1 loss integrated over Y at all $\vec{\mathbf{X}}=\vec{\mathbf{x}}$, making again the marginal distribution of $\vec{\mathbf{X}}$ irrelevant.
- Independence of the regressor distribution implies that if the joint distribution is reweighted with a function of the regressors, $\tilde{\mathbf{P}}_{Y,\vec{\mathbf{X}}} = \omega(\vec{\mathbf{X}})\mathbf{P}_{Y,\vec{\mathbf{X}}}$ where $\omega(\vec{\mathbf{X}}) \geq 0$ and $\mathbf{E}_{\mathbf{P}}[\omega(\vec{\mathbf{X}})] = 1$, then the conditional response distribution is unchanged, $\tilde{\mathbf{P}}_{Y|\vec{\mathbf{X}}} = \mathbf{P}_{Y|\vec{\mathbf{X}}}$, whereas the marginal regressor distribution absorbs the reweighting, $\tilde{\mathbf{P}}_{\vec{\mathbf{X}}} = \omega(\vec{\mathbf{X}})\mathbf{P}_{\vec{\mathbf{X}}}$. For a statistical functional $\boldsymbol{\theta}$ that is well-specified for $\mathbf{P}_{Y,\vec{\mathbf{X}}}$, we therefore have $\boldsymbol{\theta}(\tilde{\mathbf{P}}_{Y,\vec{\mathbf{X}}}) = \boldsymbol{\theta}(\mathbf{P}_{Y,\vec{\mathbf{X}}})$. This idea leads to misspecification tests based on reweighting of the data as explored by White (1980a, Section 4) for linear OLS. The idea generalizes to arbitrary statistical functionals and arbitrary data types.
- From the previous point follows that functionals that are well-specified for a particular conditional response distribution allow consistent estimation under arbitrary regressor-dependent reweighting of the data. This knowledge is implicit in much of model-trusting theory and methodology.
- The idea of regressor-dependent reweighting points toward smoothing with Parzen kernels in order to localize the statistical functional as a diagnostic for well-specification. Smoothing, of course, is a technology in its own right comprising the whole literature of nonparametric function estimation.

7. 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 

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

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

(1980b), for example, assumes observations to be “independent not (necessarily) identically distributed”, and Hansen (1982) assumes them stationary and ergodic. White’s (1994) monograph includes “dynamic misspecification” for time series aspects of the data. We limit ourselves to i.i.d. observations $(Y_i, \vec{\mathbf{X}}_i') = (Y_i, X_{i,1}, \dots, X_{i,p})$ ($i = 1, 2, \dots, N$) drawn from a multivariate distribution $\mathbf{P}(dy, dx_1, \dots, dx_p)$, and we stack them to matrices and vectors as in Table 3.

7.1 Estimation for Linear OLS

The population versions of nonlinearities η , noise ϵ , and population residuals δ translate to random N -vectors as follows (again, see Table 3):

$$(19) \quad \boldsymbol{\eta} = \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}, \quad \boldsymbol{\delta} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\eta} + \boldsymbol{\epsilon}.$$

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

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

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

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

$$(20) \quad \hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\tilde{\boldsymbol{\beta}}} \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\|^2 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

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

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

7.2 Decomposition of OLS Estimates According to Noise and Nonlinearity

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

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

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

$$(21) \quad \hat{\beta} - \beta = (\hat{\beta} - \beta(\mathbf{X})) + (\beta(\mathbf{X}) - \beta).$$

Definition and Lemma 7.2: Define “Estimation Offsets” (EOs) as follows:

$$(22) \quad \begin{array}{lll} \text{Total EO} & := \hat{\beta} - \beta & = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta, \\ \text{Noise EO} & := \hat{\beta} - \beta(\mathbf{X}) & = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon, \\ \text{Approximation EO} & := \beta(\mathbf{X}) - \beta & = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\eta. \end{array}$$

The right hand equalities follow from the decompositions (19), $\epsilon = \mathbf{Y} - \boldsymbol{\mu}$, $\eta = \boldsymbol{\mu} - \mathbf{X}\beta$, $\delta = \mathbf{Y} - \mathbf{X}\beta$, and these facts:

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

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

7.3 Generalization of the Decomposition to Statistical Functionals

The three EOs can be generalized. Starting with MoM estimators, the moment conditions that define $\boldsymbol{\theta}$, $\boldsymbol{\theta}(\mathbf{X})$ and $\hat{\boldsymbol{\theta}}$ are, respectively:

$$(23) \quad \begin{array}{lll} \boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{P}) : & \mathbf{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\mathbf{X}})] & = \mathbf{0}, \\ \boldsymbol{\theta}(\mathbf{X}) : & \frac{1}{N} \sum_i \mathbf{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y_i, \vec{\mathbf{X}}_i) | \vec{\mathbf{X}}_i] & = \mathbf{0}, \\ \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\mathbf{P}}) : & \frac{1}{N} \sum_i \boldsymbol{\psi}(\boldsymbol{\theta}; Y_i, \vec{\mathbf{X}}_i) & = \mathbf{0}. \end{array}$$

These specialize to linear OLS for $\boldsymbol{\psi}_{OLS}(\boldsymbol{\beta}; y, \vec{x})$ in (12). — The generalization to arbitrary statistical functionals is as follows, using the notation of (18):

$$(24) \quad \begin{array}{lll} \boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{P}) & = \boldsymbol{\theta}(\mathbf{P}_{Y|\vec{\mathbf{X}}}, \mathbf{P}_{\vec{\mathbf{X}}}), \\ \boldsymbol{\theta}(\mathbf{X}) & = \boldsymbol{\theta}(\mathbf{P}_{Y|\vec{\mathbf{X}}}, \hat{\mathbf{P}}_{\vec{\mathbf{X}}}), \\ \hat{\boldsymbol{\theta}} & = \boldsymbol{\theta}(\hat{\mathbf{P}}). \end{array}$$

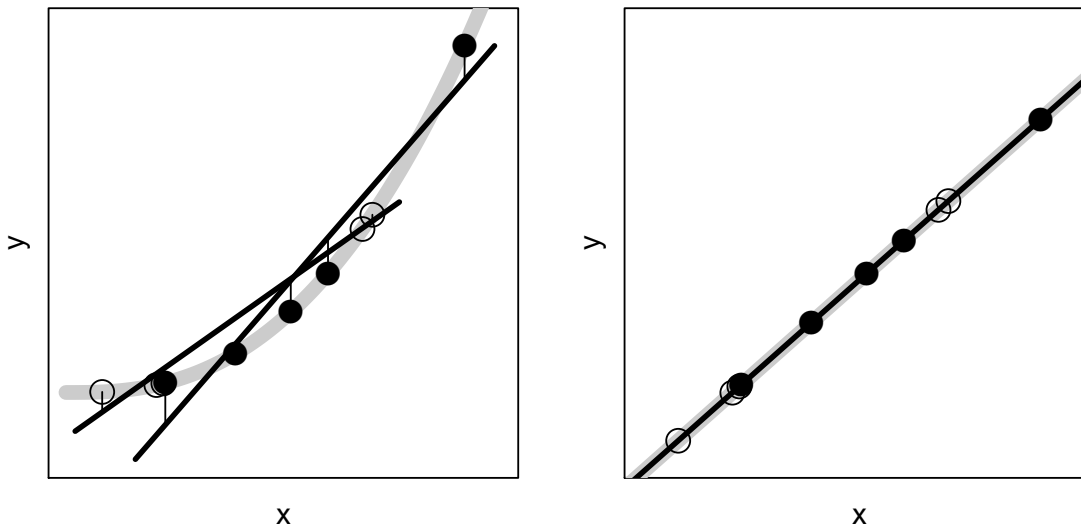


FIG 4. *Noise-less Response: The filled and the open circles represent two “datasets” from the same population. The x -values are random; the y -values are a deterministic function of x : $y = \mu(x)$ (shown in gray).*

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).

The centerpiece is the definition of $\theta(\mathbf{X})$, which is the \mathbf{X} -conditional “parameter” that integrates the conditional response distribution but conditions on the observed regressors as collected in the regressor matrix \mathbf{X} (permitting in this generality arbitrary variable types). The observed regressor data are represented by the empirical regressor distribution $\hat{\mathbf{P}}_{\vec{\mathbf{X}}} = (1/N) \sum \delta_{\vec{\mathbf{X}}_i}$ (where $\delta_{\vec{\mathbf{X}}_i}$ denotes a point mass at $\vec{\mathbf{X}}_i$, in deviation from the meaning of δ elsewhere as population residual). Estimation offsets can now be defined in full generality:

$$(25) \quad \begin{array}{ll} \text{Total EO} & := \hat{\theta} - \theta, \\ \text{Noise EO} & := \hat{\theta} - \theta(\mathbf{X}), \\ \text{Approximation EO} & := \theta(\mathbf{X}) - \theta. \end{array}$$

The target of estimation in model-trusting theories is $\theta(\mathbf{X})$, while in model-robust theories it is $\theta(\mathbf{P})$.

7.4 Random \mathbf{X} and Misspecification as a Source of Sampling Variation

The following observation further validates the definition of well-specification given in Section 6:

Lemma 7.4: *If the statistical functional θ is well-specified for $\mathbf{P}_{\mathbf{Y}|\vec{\mathbf{X}}}$, then the conditional parameter $\theta(\mathbf{X})$ agrees with the unconditional parameter $\theta(\mathbf{P})$:*

$$\theta(\mathbf{X}) \stackrel{\mathbf{P}}{=} \theta(\mathbf{P}).$$

Proof: $\theta(\mathbf{X}) = \theta(\mathbf{P}_{\mathbf{Y}|\vec{\mathbf{X}}}, \hat{\mathbf{P}}_{\vec{\mathbf{X}}}) = \theta(\mathbf{P}_{\mathbf{Y}|\vec{\mathbf{X}}}) = \theta(\mathbf{P})$ due to well-specification in the sense of Section 6. \square

If, however, there is misspecification, then $\boldsymbol{\theta}(\mathbf{X})$ becomes a random variable with genuine sampling variability. For this to occur, both randomness of \mathbf{X} and misspecification need to be present — the “conspiracy” in the title of the article. The full variability of $\hat{\boldsymbol{\theta}}$ is then no longer due to the conditional response distribution $\mathbf{P}_{Y|\vec{\mathbf{X}}}$ alone.

This fact is best illustrated with the example of a misspecified deterministic response, where $Y = \mu(\vec{\mathbf{X}})$ (that is, $\mathbf{P}_{Y|\vec{\mathbf{X}}} = \delta_{\mu(\vec{\mathbf{X}})}$ are point masses) for some non-linear function. This is shown in the left hand frame of Figure 4 for a single regressor, with OLS lines fitted to two “datasets” consisting of $N = 5$ regressor values each. The randomness in the regressors causes the fitted line to exhibit sampling variability due to the nonlinearity of the response. This effect is absent for a linear response (well-specification) shown in the right hand frame.

A comparison with Figure 2 illustrates the fact that the effect is the same in both, shown for different regressor populations in Figure 2 and for different datasets in Figure 4, corroborating the one-line proof of Lemma 7.4. Thus misspecification has two fundamental effects: (1) the population parameter $\boldsymbol{\beta}(\mathbf{P})$ becomes dependent on the regressor distribution, and (2) the “conditional parameter” $\boldsymbol{\theta}(\mathbf{X})$ exhibits sampling variability.

The effect is of course of a very general nature, not limited to linear functions fitted to nonlinear curves. It also takes place, for example, if one fits continuous functions to discontinuous response functions, or smooth functions to non-smooth response functions, or additive functions to interaction response functions.

Fixed- \mathbf{X} theories of regression that condition on the regressors, such as linear models theory, necessarily assume well-specification in the sense of Section 6. Their only source of sampling variability is the noise EO $\boldsymbol{\theta} - \boldsymbol{\theta}(\mathbf{X})$. The “remedy” of fixed- \mathbf{X} theories is to call for model diagnostics and declare a model and its inferences to be invalid if misspecification is detected. If there exist misspecifications that cause $\boldsymbol{\theta}(\mathbf{X}) - \boldsymbol{\theta} \neq \mathbf{0}$ but remain undetected in a particular data analysis, they will be erroneously treated as arising from noise, and statistical inference may consequently be invalidated (Section 9.3.4). This mistreatment occurs also in the residual bootstrap which assumes the residuals to originate from exchangeable noise. Asymptotically correct treatment is provided by the sandwich estimator and the x - y bootstrap, even in noise-free misspecified situations. The justifications derive from central limit theorems to be described next.

7.5 Model-Robust Central Limit Theorems

For a well-behaved statistical functional $\boldsymbol{\theta}(\mathbf{P})$ that has an influence function $\mathbf{IC}_{\boldsymbol{\theta},\mathbf{P}}(y, \vec{\mathbf{x}})$ (Hampel et al. 1986), the EOs have the following CLTs:

$$(26) \quad \begin{array}{l} \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{V}[\mathbf{IC}_{\boldsymbol{\theta},\mathbf{P}}(Y, \vec{\mathbf{X}})]\right), \\ \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\mathbf{X})) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{E}[\mathbf{V}[\mathbf{IC}_{\boldsymbol{\theta},\mathbf{P}}(Y, \vec{\mathbf{X}}) | \vec{\mathbf{X}}]]\right), \\ \sqrt{N}(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{V}[\mathbf{E}[\mathbf{IC}_{\boldsymbol{\theta},\mathbf{P}}(Y, \vec{\mathbf{X}}) | \vec{\mathbf{X}}]]\right). \end{array}$$

The asymptotic variances of the EOs naturally follow the identity

$$\mathbf{V}[\mathbf{IC}_{\boldsymbol{\theta},\mathbf{P}}(Y, \vec{\mathbf{X}})] = \mathbf{E}[\mathbf{V}[\mathbf{IC}_{\boldsymbol{\theta},\mathbf{P}}(Y, \vec{\mathbf{X}}) | \vec{\mathbf{X}}]] + \mathbf{V}[\mathbf{E}[\mathbf{IC}_{\boldsymbol{\theta},\mathbf{P}}(Y, \vec{\mathbf{X}}) | \vec{\mathbf{X}}]],$$

the three terms corresponding to total, noise and approximation EO, respectively.

For moment estimators the influence function can be derived from the moment condition $\mathbf{E}_P[\boldsymbol{\psi}(Y, \vec{\mathbf{X}}; \boldsymbol{\theta})] = \mathbf{0}$ (Section 3.3):

$$\mathbf{IC}_{\boldsymbol{\theta}, P}(y, \vec{\mathbf{x}}) = \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1} \boldsymbol{\psi}(\boldsymbol{\theta}; y, \vec{\mathbf{x}}),$$

where $\boldsymbol{\Lambda}(\boldsymbol{\theta}) := \partial_{\boldsymbol{\theta}} \mathbf{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\mathbf{X}})]$ is a Jacobian of size $q \times q$, $q = \dim(\boldsymbol{\psi}) = \dim(\boldsymbol{\theta})$. Asymptotic normality of the total EO is Huber's (1967) result. The asymptotic variance of the total EO has the following sandwich form:

$$(27) \quad \boxed{\mathbf{AV}[\boldsymbol{\theta}, P] = \boldsymbol{\Lambda}(\boldsymbol{\theta})^{-1} \mathbf{V}[\boldsymbol{\psi}(\boldsymbol{\theta}; Y, \vec{\mathbf{X}})] \boldsymbol{\Lambda}(\boldsymbol{\theta})'^{-1}.}$$

Linear OLS has $\mathbf{IC}_{\boldsymbol{\beta}, P}(y, \vec{\mathbf{x}}) = \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1} (\vec{\mathbf{x}} \vec{\mathbf{x}}' \boldsymbol{\beta} - \vec{\mathbf{x}} y)$, and hence the following CLTs for the EOs:

$$(28) \quad \boxed{\begin{aligned} \sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1} \mathbf{E}[m^2(\vec{\mathbf{X}}) \vec{\mathbf{X}} \vec{\mathbf{X}}'] \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1}\right) \\ \sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(\mathbf{X})) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1} \mathbf{E}[\sigma^2(\vec{\mathbf{X}}) \vec{\mathbf{X}} \vec{\mathbf{X}}'] \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1}\right) \\ \sqrt{N}(\boldsymbol{\beta}(\mathbf{X}) - \boldsymbol{\beta}) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1} \mathbf{E}[\eta^2(\vec{\mathbf{X}}) \vec{\mathbf{X}} \vec{\mathbf{X}}'] \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1}\right) \end{aligned}}$$

According to (15) and (16), $m^2(\vec{\mathbf{X}})$ can be replaced by δ^2 and $\sigma^2(\vec{\mathbf{X}})$ by ϵ^2 :

$$(29) \quad \mathbf{E}[m^2(\vec{\mathbf{X}}) \vec{\mathbf{X}} \vec{\mathbf{X}}'] = \mathbf{E}[\delta^2 \vec{\mathbf{X}} \vec{\mathbf{X}}'], \quad \mathbf{E}[\sigma^2(\vec{\mathbf{X}}) \vec{\mathbf{X}} \vec{\mathbf{X}}'] = \mathbf{E}[\epsilon^2 \vec{\mathbf{X}} \vec{\mathbf{X}}'].$$

The asymptotic variance of linear OLS can therefore be written as

$$(30) \quad \boxed{\mathbf{AV}[\boldsymbol{\beta}, P] = \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1} \mathbf{E}[\delta^2 \vec{\mathbf{X}} \vec{\mathbf{X}}'] \mathbf{E}[\vec{\mathbf{X}} \vec{\mathbf{X}}']^{-1}.}$$

Special cases covered by these CLTs are as follows:

- **First order well-specification:** $\eta(\vec{\mathbf{X}}) \stackrel{P}{=} 0$. The sandwich form is solely due to heteroskedasticity.
- **Deterministic nonlinear response:** $\sigma^2(\vec{\mathbf{X}}) \stackrel{P}{=} 0$. The sandwich form is solely due to the nonlinearity and randomness of \mathbf{X} .
- **First and second order well-specification:** $\eta(\vec{\mathbf{X}}) \stackrel{P}{=} 0$, $\sigma^2(\vec{\mathbf{X}}) \stackrel{P}{=} \sigma^2(P)$. The non-sandwich form is asymptotically valid without Gaussian errors.

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

Empirically one observes that standard error estimates obtained from the x - y bootstrap and from the sandwich estimator are generally close to each other. This is intuitively unsurprising as they both estimate the same asymptotic variance, that of the first CLT in Proposition 7.5. A closer connection between them will be established here.

8.1 The Plug-In Sandwich Estimator of Asymptotic Variance

Plug-in estimators of standard error are obtained by substituting the empirical distribution \hat{P} for the true P in formulas for asymptotic variances (26):

$$(31) \quad \hat{AV}[\hat{\theta}] = AV[\theta, \hat{P}], \quad \hat{SE}[\hat{\theta}_j] := \frac{1}{N^{1/2}}(\hat{AV}[\hat{\theta}])_{jj}^{1/2}.$$

These are asymptotically consistent under mild conditions.

For MoM-estimators (31) specializes to the following sandwich form, using (27) and writing $\hat{E}[\dots]$ and $\hat{V}[\dots]$ for sample means and sample variance/covariances over (Y_i, \vec{X}_i) ($i = 1, \dots, N$):

$$(32) \quad \hat{AV}[\hat{\theta}] := \hat{\Lambda}^{-1} \hat{V}[\psi(\hat{\theta}; Y, \vec{X})] \hat{\Lambda}'^{-1}, \quad \text{where } \hat{\Lambda} := \hat{E}[\partial_{\theta} \psi(\hat{\theta}; Y, \vec{X})].$$

The OLS sandwich estimator is the plug-in version of (30) where δ^2 is replaced by residuals, summarized in a diagonal matrix $D(\mathbf{r})^2$ with squared residuals $r_i^2 = (Y_i - \vec{X}_i \hat{\beta})^2$ in the diagonal, and

$$\hat{E}[\vec{X} \vec{X}'] = \frac{1}{N} (\mathbf{X}'\mathbf{X}), \quad \hat{E}[r^2 \vec{X} \vec{X}'] = \frac{1}{N} (\mathbf{X}'D(\mathbf{r})^2 \mathbf{X}).$$

The original sandwich estimator for linear OLS (White 1980a) can be written as

$$(33) \quad \begin{aligned} \hat{AV}_{sand}[\hat{\beta}] &:= \hat{E}[\vec{X} \vec{X}']^{-1} \hat{E}[r^2 \vec{X} \vec{X}'] \hat{E}[\vec{X} \vec{X}']^{-1} \\ &= N (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'D(\mathbf{r})^2 \mathbf{X}) (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

This is version ‘‘HC’’ in MacKinnon and White (1985). A modification accounts for the fact that residuals have smaller variance than noise, calling for a correction by replacing $1/N^{1/2}$ in (31) with $1/(N - p - 1)^{1/2}$, in analogy to the linear models estimator (‘‘HC1’’ *ibid.*). Another modification is to correct individual residuals for their reduced variance according to $V[r_i | \mathbf{X}] = \sigma^2(1 - H_{ii})$ under homoskedasticity and ignoring nonlinearity (‘‘HC2’’ *ibid.*). Further modifications include a version based on the jackknife (‘‘HC3’’ *ibid.*) using leave-one-out residuals.

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

An alternative to plug-in is estimating asymptotic variance with the x - y bootstrap. To link plug-in and bootstrap estimators we need the M -of- N bootstrap where the *resample size* M may differ from the sample size N . One distinguishes

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

In resampling, M can be any $M < \infty$; in subsampling, M must satisfy $M < N$. The M -of- N bootstrap for $M \ll N$ ‘‘works’’ more often than the conventional N -of- N bootstrap; see Bickel, Götze and van Zwet (1997) who showed that the favorable properties of $M \ll N$ subsampling obtained by Politis and Romano (1994) carry over to the $M \ll N$ bootstrap. Ours is a well behaved context, hence there is no need for $M \ll N$; instead, we consider bootstrap resampling for the extreme case $M \gg N$, namely, the limit $M \rightarrow \infty$.

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, $M \rightarrow \infty$. It is immaterial that, in this case, the sampled distribution is the empirical

distribution $\hat{\mathbf{P}} = \hat{\mathbf{P}}_N$ of a given dataset $\{(Y_i, \vec{\mathbf{X}}_i)\}_{i=1\dots N}$, which is frozen of size N as $M \rightarrow \infty$. The following holds for bootstrap resampling of any well-behaved statistical functional, be it in a regression context or not:

Proposition 8.2: *For any fixed dataset of size N , represented by $\hat{\mathbf{P}}_N$, if $\boldsymbol{\theta}$ is asymptotically normal, there holds a CLT for the M -of- N bootstrap as $M \rightarrow \infty$, with an asymptotic variance obtained by plug-in. Letting $\boldsymbol{\theta}^* = \boldsymbol{\theta}(\mathbf{P}^*)$ where \mathbf{P}^* represents a resample of size M from $\hat{\mathbf{P}}_N$, we have:*

$$(34) \quad M^{1/2}(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{AV}[\boldsymbol{\theta}, \hat{\mathbf{P}}_N]\right) \quad (M \rightarrow \infty, N \text{ fixed}).$$

No proof is needed. We immediately have the following:

Corollary 8.2: *The sandwich estimators (32) and (33) for MoM and OLS estimators are the asymptotic variance estimated by the M -of- N bootstrap in the limit $M \rightarrow \infty$ for a fixed sample of size N .*

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

For further connections see MacKinnon and White (1985): Some forms of sandwich estimators were independently derived by Efron (1982, p. 18f) using the infinitesimal jackknife, and by Hinkley (1977) using a “weighted jackknife.” See Weber (1986) for a concise comparison in the linear model limited to heteroskedasticity. A deep connection between jackknife and bootstrap is given by Wu (1986).

9. INSIGHTS FROM LINEAR OLS

The following long section will use the special structure of OLS to make points that are generalizable in principle but best explained in a simple context: the meaning of regression slopes in the presence of nonlinearity (= misspecification for slopes), an asymptotic comparison of model-trusting and model-robust standard errors, scenarios of misspecification that affect the ratio of the two kinds of asymptotic variances, and finally a test for the discrepancy between them.

9.1 Adjusted Regressors

The following adjustment formulas are standard but will be stated explicitly due to their importance in what follows. They express the slopes of multiple regressions as slopes of simple regressions using adjusted single regressors. The formulas will be used for the interpretation of regression slopes in the presence of nonlinearity (Section 9.2), the analysis of discrepancies between asymptotically proper and improper standard errors (Section 9.3), and a test of discrepancy between the two (Section 9.4). [See Appendix C for more notational details.]

- **Adjustment in Populations:** The population-adjusted regressor random variable $X_{j\bullet}$ is the “residual” of the population regression of X_j , used as the response, on all other regressors. The response Y can be adjusted similarly, and we may denote it by $Y_{\bullet-j}$ to indicate that X_j is not among the adjustors, which is implicit in the adjustment of X_j . The multiple regression

coefficient $\beta_j = \beta_j(\mathbf{P})$ of the population regression of Y on $\vec{\mathbf{X}}$ is obtained as the simple regression through the origin of Y or $Y_{\bullet-j}$ on $X_{j\bullet}$:

$$(35) \quad \beta_j = \frac{E[Y_{\bullet-j}X_{j\bullet}]}{E[X_{j\bullet}^2]} = \frac{E[YX_{j\bullet}]}{E[X_{j\bullet}^2]} = \frac{E[\mu(\vec{\mathbf{X}})X_{j\bullet}]}{E[X_{j\bullet}^2]}.$$

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

- **Adjustment in Samples:** Define the sample-adjusted regressor column $\mathbf{X}_{j\hat{\bullet}}$ to be the residual vector of the sample regression of \mathbf{X}_j , used as the response vector, on all other regressors. The response vector \mathbf{Y} can be sample-adjusted similarly, and we may denote it by $\mathbf{Y}_{\bullet-j}$ to indicate that \mathbf{X}_j is not among the adjustors, which is implicit for $X_{j\bullet}$. (Note the use of hat notation “ $\hat{\bullet}$ ” to distinguish it from population-based adjustment “ \bullet .”) The coefficient estimate $\hat{\beta}_j$ of the multiple regression of \mathbf{Y} on \mathbf{X} is obtained as the simple regression through the origin of \mathbf{Y} or $\mathbf{Y}_{\bullet-j}$ on $\mathbf{X}_{j\hat{\bullet}}$:

$$(36) \quad \hat{\beta}_j = \frac{\langle \mathbf{Y}_{\bullet-j}, \mathbf{X}_{j\hat{\bullet}} \rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^2} = \frac{\langle \mathbf{Y}, \mathbf{X}_{j\hat{\bullet}} \rangle}{\|\mathbf{X}_{j\hat{\bullet}}\|^2}.$$

9.2 The Meaning of Slopes in the Presence of Nonlinearity

A first use of regressor adjustment is for proposing a meaning of linear slopes in the presence of nonlinearity, and thereby responding to Freedman’s (2006, p. 302) objection: “... it is quite another thing to ignore bias [nonlinearity]. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter.” Against this view one may hold that the parameter is not intrinsically wrong, rather, it is in need of a useful interpretation: a linear fit in the presence of nonlinearity gives a sense of the direction, up or down, of association between a regressor and the response adjusted for other regressors. (If the sole purpose is response prediction, well-specification is not the goal either; it is rather trading off nonlinearity against noise over the regressor range.)

The issue is that, in the presence of nonlinearity, slopes lose their usual interpretation: β_j is no longer the average difference in Y associated with a unit difference in X_j at fixed levels of all other X_k . The challenge is to provide an alternative interpretation that remains valid and intuitive. As mentioned, a plausible approach is to use adjusted variables, in which case it is sufficient to solve the interpretation problem for simple regression through the origin. Regression slopes can then be interpreted as weighted averages of “case-wise” and “pairwise” slopes in a sense to be made precise. This interpretation holds even for regressors that are nonlinearly related, as in $X_2 = X_1^2$ or $X_3 = X_1X_2$, because the clause “at fixed levels of all other regressors” is replaced by reference to “(linearly) adjusted regressors.” (“Linearly” will be dropped in what follows.)

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

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

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

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

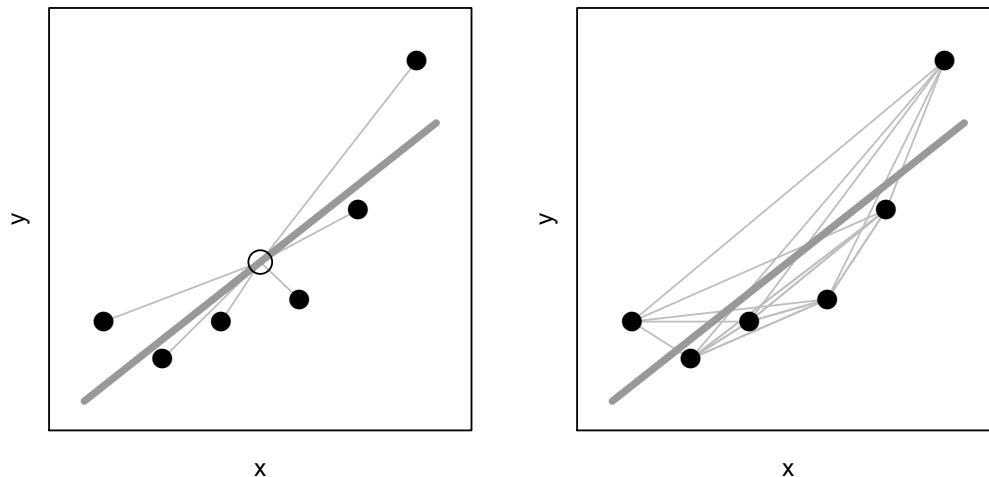


FIG 5. *Case-wise and pairwise average weighted slopes illustrated: Both plots show the same six points (“cases”) as well as the OLS line fitted to them (fat gray). The left hand plot shows the case-wise slopes from the mean point (open circle) to the six cases, while the right hand plot shows the pairwise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

Slope interpretation will be based on the following devices:

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

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

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

- **pairwise slopes:**

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

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

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

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

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

- **pairwise slopes:**

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

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

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

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

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

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

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

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

Note on literature: The above formulas were used and modified to produce alternative slope estimates by Gelman and Park (2008), with the “Goal of Expressing Regressions as Comparisons that can be Understood by the General Reader” (see their Sections 1.2 and 2.2). Earlier, Wu (1986) used generalizations based on tuples rather than pairs of (y_i, \vec{x}'_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.

9.3 Asymptotic Variances — Proper and Improper

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

9.3.1 Preliminaries: Adjustment Formulas for EOs and Their CLTs: The vectorized formulas for estimation offsets (21) can be written componentwise using adjustment as follows:

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

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

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

Corollary 9.3.1:

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

The equalities on the right side in the first and second case are based on (29). The first one will be needed for plug-in estimation. The sandwich form for matrices has been reduced to a ratio where the numerator corresponds to the “meat” and the squared denominator to the “breads”.

9.3.2 Model-Robust Asymptotic Variances in Terms of Adjusted Regressors: The CLTs of Corollary 9.3.1 contain three asymptotic variances of the same form with arguments $m^2(\vec{\mathbf{X}})$, $\sigma^2(\vec{\mathbf{X}})$ and $\eta^2(\vec{\mathbf{X}})$. This suggests using generic notation:

Definition 9.3.2: *Proper Asymptotic Variance and its Components.*

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

The subscript *lean* refers to the assumption-lean model-robust framework.

9.3.3 Model-Trusting Asymptotic Variances in Terms of Adjusted Regressors: The goal is to provide an asymptotic limit for the usual model-trusting standard error estimate of linear models theory in the model-robust framework. It derives from an estimate $\hat{\sigma}^2$ of the noise variance, $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 / (N - p - 1)$, which has the following limit for fixed p :

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

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

$$(38) \quad \hat{\mathbf{V}}_{in}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}, \quad \hat{\mathbf{S}}\mathbf{E}_{in}^2[\hat{\beta}_j] = \frac{\hat{\sigma}^2}{\|\mathbf{X}_{j\bullet}\|^2}.$$

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

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

These limits are **model-trusting** asymptotic variances which provide valid standard errors if the first and second order assumptions of linear models theory hold. Here is again a generic definition with an associated decomposition:

Definition 9.3.3: *Improper Asymptotic Variance and its Components.*

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

The subscript *lin* refers to the assumption-loaded model-trusting framework of linear models theory.

9.3.4 **RAV** — *Ratio of Proper and Improper Asymptotic Variances:* To examine the discrepancies between proper and improper asymptotic variances we form their ratios separately for each of $m^2(\vec{X})$, $\sigma^2(\vec{X})$ and $\eta^2(\vec{X})$, hence we use again a generic form of the ratio:

Definition 9.3.4: *Ratio of Asymptotic Variances, Proper/Improper.*

For $f^2(\vec{x}) = m^2(\vec{x})$, $\sigma^2(\vec{x})$ or $\eta^2(\vec{x})$, let

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

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

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

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

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

Lemma 9.3.4: *If δ^2 and $X_{j\bullet}^2$ are independent, then $\mathbf{RAV}[\hat{\beta}_j, m^2] = 1$.*

Proof: The numerator of $\mathbf{RAV}[\hat{\beta}_j, m^2]$, which is $\mathbf{E}[m^2(\vec{X})X_{j\bullet}^2]$, factorizes and cancels out with the denominators.

The three terms $\mathbf{RAV}[\hat{\beta}_j, m^2]$, $\mathbf{RAV}[\hat{\beta}_j, \sigma^2]$ and $\mathbf{RAV}[\hat{\beta}_j, \eta^2]$ can be interpreted as inner products between the three random variables

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

These are *not* correlations, and they are not upper bounded by +1; their natural bounds are rather 0 and ∞ (Section 9.3.5).

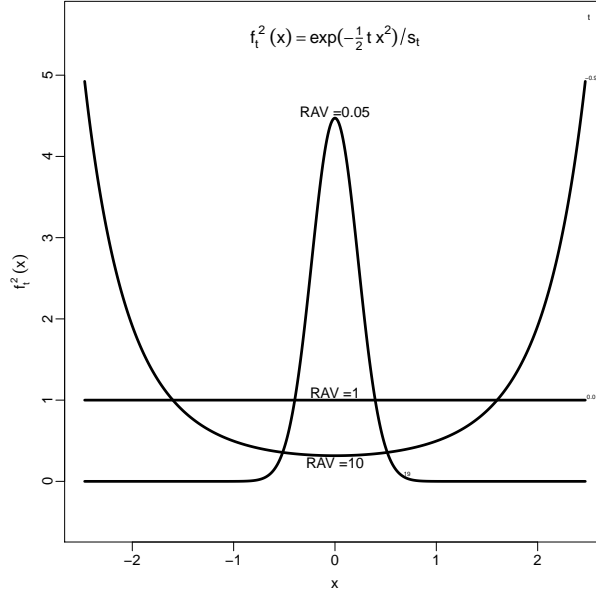


FIG 6. A family of functions $f_t^2(x)$ that can be interpreted as heteroskedasticities $\sigma_j^2(X_{j\bullet})$, squared nonlinearities $\eta_j^2(X_{j\bullet})$, or conditional MSEs $m_j^2(X_{j\bullet})$: The family interpolates \mathbf{RAV} from 0 to ∞ for $x = X_{j\bullet} \sim N(0, 1)$. The three solid black curves show $f_t^2(x)$ that result in $\mathbf{RAV}=0.05, 1$, and 10. (See Appendix D.4 for details.)

$\mathbf{RAV} = \infty$ is approached as $f_t^2(x)$ bends ever more strongly in the tails of the x -distribution. $\mathbf{RAV} = 0$ is approached by an ever stronger spike in the center of the x -distribution.

A simplification is achieved by conditioning the three left hand terms on $X_{j\bullet}^2$:

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

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

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

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

Proposition 9.3.5: If $\mathbf{E}[X_{j\bullet}^2] < \infty$ and $X_{j\bullet}$ has unbounded support, then

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

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

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

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

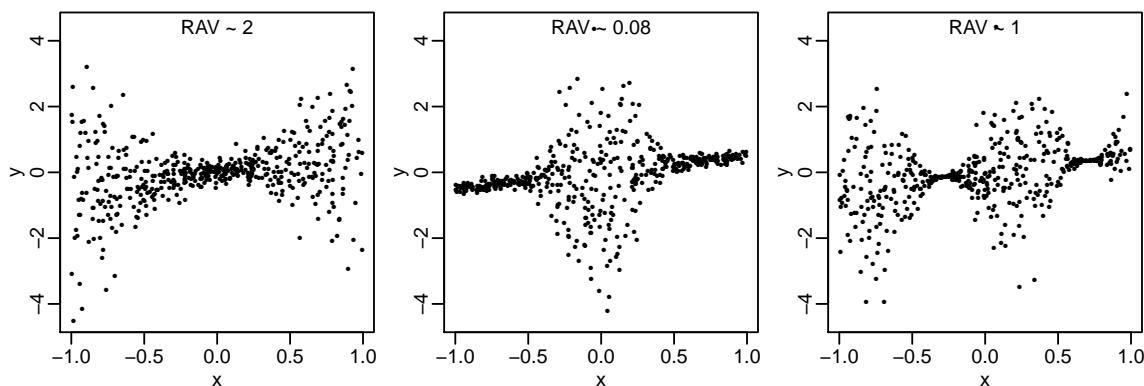


FIG 7. *The effect of heteroskedasticity on the sampling variability of slope estimates: How does the treatment of the heteroskedasticities as homoskedastic affect statistical inference?*
 Left: *High noise variance in the tails of the regressor distribution elevates the true sampling variability of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

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

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

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

9.3.6 Illustration of Factors that Drive the RAV: To further analyze the \mathbf{RAV} , we drill down from $m_j^2(X_{j\bullet}^2)$ to $\sigma_j^2(X_{j\bullet}^2)$ and $\eta_j^2(X_{j\bullet}^2)$ 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 \mathbf{RAV} . According to the \mathbf{RAV} decomposition lemma in Appendix D.2, $\mathbf{RAV}[\hat{\beta}_j, m_j^2]$ is a mixture of $\mathbf{RAV}[\hat{\beta}_j, \sigma_j^2]$ and $\mathbf{RAV}[\hat{\beta}_j, \eta_j^2]$. Therefore:

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

These facts also suggest that large values $\mathbf{RAV} > 1$ should occur more often than small values $\mathbf{RAV} < 1$ because large conditional variances as well as nonlinearities are often more pronounced in the extremes of regressor distributions, not their centers. This is most natural for nonlinearities which are often convex or concave.

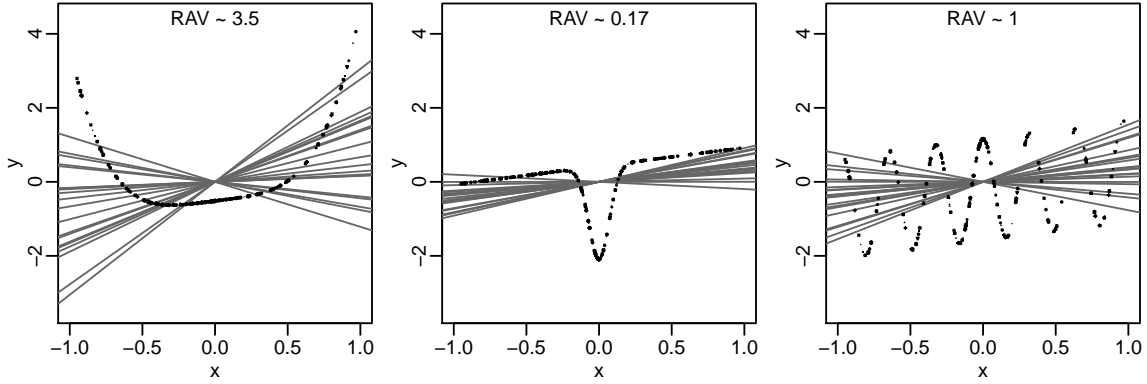


FIG 8. *The effect of nonlinearities on the sampling variability of slope estimates: The three plots show three different noise-free nonlinearities; each plot shows for one nonlinearity 20 overplotted datasets of size $N = 10$ and their fitted lines through the origin. The question is how the misinterpretation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

Also, it follows from the \mathbf{RAV} decomposition lemma (Appendix D.2) that either of $\mathbf{RAV}[\hat{\beta}_j, \sigma_j^2]$ or $\mathbf{RAV}[\hat{\beta}_j, \eta_j^2]$ is able to single-handedly pull $\mathbf{RAV}[\hat{\beta}_j, m_j^2]$ to $+\infty$, whereas both have to be close to zero to pull $\mathbf{RAV}[\hat{\beta}_j, m_j^2]$ toward zero. These considerations are heuristics for the observation that in practice $\hat{\mathbf{SE}}_{lin}$ is more often too small than too large compared to $\hat{\mathbf{SE}}_{sand}$.

9.4 Sandwich Estimators in Adjusted Form and a \mathbf{RAV} Test

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

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

As illustrated above, the types of nonlinearities and heteroskedasticities that result in discrepancies between \mathbf{SE}_{lin} and \mathbf{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.

9.4.1 Sandwich Estimators in Adjustment Form and the \mathbf{RAV}_j Test Statistic: To begin with, the adjustment versions of the asymptotic variances in the CLTs of Corollary 9.3.1 can be used to rewrite the sandwich estimator by replacing expectations $\mathbf{E}[\dots]$ with means $\hat{\mathbf{E}}[\dots]$, β with $\hat{\beta}$, $X_{j\bullet}$ with $\mathbf{X}_{j\bullet}$, and rescaling by N :

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

The squaring of N -vectors is meant to be coordinate-wise. Formula (40) is algebraically equivalent to the diagonal elements of (33).

The usual squared standard error estimate (38) is

$$(41) \quad \hat{\mathbf{SE}}_{lin}[\hat{\beta}_j]^2 = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2}{(N-p-1)\|\mathbf{X}_{j\bullet}\|^2} \sim \frac{1}{N} \frac{\hat{\mathbf{E}}[(Y - \bar{\mathbf{X}}'\hat{\beta})^2]}{\hat{\mathbf{E}}[X_{j\bullet}^2]} = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2}{N\|\mathbf{X}_{j\bullet}\|^2},$$

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

$$(42) \quad \mathbf{RAV}_j := \frac{\hat{\mathbf{E}}[(Y - \bar{\mathbf{X}}'\hat{\beta})^2 X_{j\bullet}^2]}{\hat{\mathbf{E}}[(Y - \bar{\mathbf{X}}'\hat{\beta})^2] \hat{\mathbf{E}}[X_{j\bullet}^2]} = N \frac{\langle (Y - \mathbf{X}\hat{\beta})^2, \mathbf{X}_{j\bullet}^2 \rangle}{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 \|\mathbf{X}_{j\bullet}\|^2}.$$

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

9.4.2 The Asymptotic Null Distribution of the \mathbf{RAV} Test Statistic: Here is an asymptotic result that would be expected to yield approximate inference under a null hypothesis that implies $\mathbf{RAV}[\hat{\beta}_j, m^2] = 1$ based on Lemma 9.3.4:

Proposition 9.4.2: *Under the null hypothesis H_0 that the population residuals δ and the adjusted regressor $X_{j\bullet}$ are independent, it holds:*

$$(43) \quad N^{1/2} (\mathbf{RAV}_j - 1) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, \frac{\mathbf{E}[\delta^4]}{\mathbf{E}[\delta^2]^2} \frac{\mathbf{E}[X_{j\bullet}^4]}{\mathbf{E}[X_{j\bullet}^2]^2} - 1 \right).$$

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

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

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

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

TABLE 4

LA Homeless data: Permutation Inference for \hat{RAV}_j (10,000 permutations). Values of \hat{RAV}_j that fall outside the middle 95% range of their permutation null distributions are marked with asterisks. They indicate statistically significant deviations of the usual model-trusting standard errors of linear models theory from their model-robust sandwich analogs. For MedianInc (1000) and PercResidential the usual standard error is too large (conservative), while for PercIndustrial it is too small (liberal). Surprisingly, the values of approximately 2 for the \hat{RAV}_j of PercVacant and PercCommercial are not statistically significant.

9.4.3 An Approximate Permutation Distribution of the \mathbf{RAV} Test Statistic:

The asymptotic result of Proposition 9.4.2 provides qualitative insights, but it is not suitable for practical application because the null distribution of \mathbf{RAV}_j can be very non-normal for finite N , and this in ways that are not easily overcome with simple tools such as nonlinear transformations. Another approach to null distributions for finite N is needed, and it is available in the form of an approximate permutation test because H_0 is just a null hypothesis of independence, here between δ and $X_{j\bullet}$. The test is not exact, requiring $N \gg p$, because population residuals δ_i must be estimated with sample residuals r_i and population adjusted regressor values $X_{i,j\bullet}$ with sample adjusted analogs $X_{i,j\bullet}$. The permutation simulation is cheap: Once coordinate-wise squared vectors \mathbf{r}^2 and $\mathbf{X}_{j\bullet}^2$ are formed, a draw from the conditional null distribution of \mathbf{RAV}_j is obtained by randomly permuting one of the vectors and forming the inner product with the other vector, rescaled by a fixed factor $N/(\|\mathbf{r}\|^2\|\mathbf{X}_{j\bullet}\|^2)$. A retention interval should be formed directly from the $\alpha/2$ and $1-\alpha/2$ quantiles of the permutation distribution to account for distributional asymmetries. Additionally, the permutation distribution yields an easy diagnostic of non-normality (see Appendix E for examples). — Table 4 illustrates \mathbf{RAV} tests with the LA Homeless data.

9.4.4 Generalizations of \mathbf{RAV} Tests: The \mathbf{RAV} test proposed here seems to be novel. It is not a special case of White’s (1980b) global heteroskedasticity test, nor of his misspecification test for general ML estimation (White 1982). The latter is based on equating the two forms of the information matrix, hence works on the matrix-inverse scale of asymptotic variances and is incapable of comparing model-trusting and model-robust asymptotic variances of specific parameters. Generalized \mathbf{RAV} tests are conceivable for general MoM estimators by forming ratios $\hat{\mathbf{A}}_{jj}/(\hat{\mathbf{\Lambda}}^{-1})_{jj}$ using notation of Sections 7.5 (27) and 8.1 (32). We do not have results for \mathbf{RAV} tests in this generality, however.

10. ISSUES WITH MODEL-ROBUST STANDARD ERRORS

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

estimators should be accurate only when the sample size is sufficiently large. This fact suggests that use of a model-trusting standard error should be kept in mind if there is evidence in its favor, for example, through the *RAV* test (Section 9.4).

Another cost associated with the sandwich estimator is **non-robustness in the sense of robust statistics** (Huber and Ronchetti 2009, Hampel et al. 1986), meaning strong sensitivity to outlying observations and heavy-tailed error distributions: The statistic $\hat{SE}_{sand}^2[\hat{\beta}_j]$ (40) is a ratio of fourth order quantities of the data, whereas $\hat{SE}_{lin}^2[\hat{\beta}_j]$ (41) is “only” a ratio of second order quantities. [Note we are here concerned with non-robustness of standard error estimates, not parameter estimates.] It appears, therefore, that the two types of robustness are in conflict: Model-robust standard error estimators are highly non-robust compared to their model-trusting analogs. This is a large issue which we can only raise but not solve in this space. Here are a few observations and suggestions:

- If model-robust standard errors are not classically robust, anecdotal evidence indicates the converse: the standard errors of classical robust regression are not model-robust either. In the LA Homeless data, for example, for the most important variable *PercVacant*, we observed a ratio of 1:3.28 when comparing the standard error reported by the software (function *r1m* in the *R Language* (2008)) and its model-robust analog from the *x-y*-bootstrap.
- Yet classical robust regression may confer partial robustness to the sandwich standard error as it caps residuals with a bounded ψ function. This addresses robustness to outlyingness in the vertical (y) direction.
- Robustness to outlyingness in the horizontal (\vec{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 D.3 in the appendix, one might search for transformations that obviate the need for a model-robust standard error in the first place.

not sure this is the "converse"

but if model is misspecified, transforming x will affect definition of the pop'n linear slopes

To illustrate the last point, we transformed the regressors of the LA Homeless data with their empirical cdfs to achieve approximately uniform marginal distributions. The transformed data are no longer i.i.d., but the point is to examine the effect of transforming the regressors to a finite range. As a result, shown in Table 5, the discrepancies between sandwich and usual standard errors have all but disappeared. The same drastic effect is not seen in the Boston Housing data (Appendix A, Table 7), although the discrepancies are greatly reduced here, too.

11. SUMMARY AND OUTLOOK

If statistical models imply “simplification and idealization” (Cox 1995), they should be treated as approximations rather than well-specified truths. The implications of this view are vast: (1) Parameters need to be re-interpreted as statistical functionals defined on large nonparametric sets of data distributions beyond a chosen model; (2) a main function of models is to supply objective functions and moment conditions to construct such statistical functionals; (3) a notion of well/mis-specification can be defined for general statistical functionals; (4) for

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

TABLE 5

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

the same data distribution some statistical functionals may be well-specified while others may not; (5) the degree to which statistical functionals are misspecified determines the degree to which regressor ancillarity is violated; as a consequence of regressor non-ancillarity, (6) parameters (statistical functionals) depend on the regressor distribution, and (7) a source of sampling variability in estimation arises from an interaction (“conspiracy”) of regressor randomness and misspecification; (8) this sampling variability is asymptotically correctly captured by model-robust standard error estimates from the x - y bootstrap and asymptotic plug-in (which includes sandwich estimators); (9) plug-in (and hence sandwich) estimators are a limiting case of x - y bootstrap standard errors. These facts hold universally for all regression methods based on i.i.d. data and well-behaved statistical functionals.

For linear OLS we identified the nature of misspecifications that render model-trusting standard errors too optimistic or too pessimistic, or neither. If the latter, the misspecification is benign. This suggests that general-purpose model diagnostics and misspecification tests cannot be used to invalidate model-trusting standard errors. Rather, a specific test such as the **RAV** test is needed.

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

In spite of these advances, in finite data not all possibilities of misspecification can be approached simultaneously, and there arises a need for model-robust inference. Even if complex modeling is possible, simple questions may call for simple models, in which case again one may want to look for model-robust inference.

There exist, finally, areas of statistics research where model-trusting theory appears frequently:

- Bayes inference, when it relies on uninformative priors, is asymptotically

equivalent to model-trusting frequentist inference. It should be reasonable to ask how far inferences from Bayesian models are adversely affected by misspecification. Complex Bayesian models often use large numbers of fitted parameters and control overfitting by shrinkage, hence asymptotic comparisons may be inadequate and might have to be replaced by other forms of analysis. Some promising developments are the following: Szpiro, Rice and Lumley (2010) derive a sandwich estimator from Bayesian assumptions, and a lively discussion of misspecification from a Bayesian perspective involved Walker (2013), De Blasi (2013), Hoff and Wakefield (2013) and O’Hagan (2013), who provide further references.

- High-dimensional inference is the subject of a large literature that often appears to rely on the assumptions of linearity, homoskedasticity as well as normality of error distributions. It may be uncertain whether procedures proposed in this area are model-robust. Recently, however, attention to the issue started to be paid by Bühlmann and van de Geer (2015). Related is also the incorporation of ideas from robust statistics by, for example, El Karoui et al. (2013), Donoho and Montanari (2014), and Loh (2015).

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

Acknowledgments: We are grateful to Gemma Moran and Bikram Karmakar for their help in the generalizations of Section 7.

REFERENCES

- [1] ALDRICH, J. (2005). Fisher and Regression. *Statistical Science* **20** (4), 4001–417.
- [2] BERK, R., BROWN, L., BUJA, A., ZHANG, K., AND ZHAO, L. (2013). Valid Post-Selection Inference. *The Annals of Statistics* **41** (2), 802–837.
- [3] BERK, R. H. (1966). Limiting Behavior of Posterior Distributions When the Model is Incorrect. *The Annals of Mathematical Statistics* **37** (1), 51–58.
- [4] BERK, R. H. (1970). Consistency A Posteriori. *The Annals of Mathematical Statistics* **41** (3), 894–960.
- [5] BERK, R. H. and KRIEGLER, B. and YILVISAKER, D. (2008). Counting the Homeless in Los Angeles County. in *Probability and Statistics: Essays in Honor of David A. Freedman*, Monograph Series for the Institute of Mathematical Statistics, D. Nolan and S. Speed (eds.)
- [6] BERMAN, M. (1988). A Theorem of Jacobi and its Generalization. *Biometrika* **75** (4), 779–783.
- [7] BICKEL, P. J. and GÖTZE, F. and VAN ZWET, W. R. (1997). Resampling Fewer than n Observations: Gains, Losses, and Remedies for Losses. *Statistica Sinica* **7**, 1–31.
- [8] BOX, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. in *Robustness in Statistics: Proceedings of a Workshop* (Launer, R. L., and Wilkinson, G. N., eds.) Amsterdam: Academic Press (Elsevier), 201–236.
- [9] BÜHLMANN, P. and VAN DE GEER, S. (2015). High-dimensional Inference in Misspecified Linear Models. **arXiv:1503.06426**
- [10] COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*, London: Chapman & Hall.
- [11] COX, D.R. (1995). Discussion of Chatfield (1995). *Journal of the Royal Statistical Society, Series A* **158** (3), 455–456.

- [12] DAVIES, P. L. (2014). *Data Analysis and Approximate Models*. Boca Raton, FL: CRC Press.
- [13] DE BLASI, P. (2013). Discussion of Walker (2013). *Journal of Statistical Planning and Inference* **143**, 1634–1637.
- [14] DIGGLE, P. J. and HEAGERTY, P. and LIANG, K. Y., and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*. Oxford Statistical Science Series. Oxford: Oxford University Press. ISBN 978-0-19-852484-7.
- [15] DONOHO, D. D. and MONTANARI, A. (2014). Variance Breakdown of Huber (M)-estimators: $n/p \rightarrow m \in (1, \infty)$. **arXiv:1503.02106**
- [16] EFRON, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
- [17] EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press.
- [18] EICKER, F. (1963). Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions. *The Annals of Mathematical Statistics* **34** (2), 447-456.
- [19] EL KAROUI, N. and BEAN, D. and BICKEL, P. and YU, B. (2013). Optimal M-Estimation in High-Dimensional Regression. *Proceedings of National Academy of Sciences* **110** (36), 14563-14568.
- [20] FREEDMAN, D. A. (1981). Bootstrapping Regression Models. *The Annals of Statistics* **9** (6), 1218–1228.
- [21] FREEDMAN, D. A. (2006). On the So-Called “Huber Sandwich Estimator” and “Robust Standard Errors.” *The American Statistician* **60** (4), 299–302.
- [22] GELMAN, A. and PARK, D.. K. (2008). Splitting a Regressor at the Upper Quarter or Third and the Lower Quarter or Third, *The American Statistician* **62** (4), 1–8.
- [23] HALL, A. R. (2005). *Generalized Method of Moments* (Advanced Texts in Econometrics). Oxford: Oxford University Press. ISBN 0-19-877520-2.
- [24] HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. (Springer Series in Statistics) New York, NY: Springer Verlag.
- [25] HAMPEL, F. R. and RONCHETTI, E. M. and ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach based on Influence Functions*. New York, NY: Wiley.
- [26] HANSEN, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **50** (4), 10291054.
- [27] HARRISON, X. and RUBINFELD, X. (1978). Hedonic Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management* **5**, 81–102.
- [28] HAUSMAN, J. A. (1978). Specification Tests in Econometrics. *Econometrica* **46** (6), 1251-1271.
- [29] HINKLEY, D. V. (1977). Jackknifing in Unbalanced Situations. *Technometrics* **19**, 285–292.
- [30] HOFF, P. and WAKEFIELD, J. (2013). Bayesian Sandwich Posteriors for Pseudo-True Parameters — Discussion of Walker (2013). *Journal of Statistical Planning and Inference* **143**, 1638–1642.
- [31] HUBER, P. J. (1967). The Behavior of Maximum Likelihood Estimation under Nonstandard Conditions. PROCEEDINGS OF THE FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, Vol. 1, Berkeley: University of California Press, 221–233.
- [32] HUBER, P. J. and RONCHETTI, E.M. (2009). *Robust Statistics.*, 2nd ed. New York, NY: Wiley.
- [33] KAUERMANN, G. and CARROLL, R. J. (2001). A Note on the Efficiency of Sandwich Covariance Matrix Estimation, *Journal of the American Statistical Association* **96**(456), 1387-1396.
- [34] KENT, J. (1982). Robust Properties of Likelihood Ratio Tests. *Biometrika* **69** (1), 19–27.
- [35] KRASKER, W. S. and WELSCH, R. W. (1982). Efficient Bounded-Influence Regression Estimation. *Journal of the American Statistical Association* **77** (379), 595-604.
- [36] LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73** (1), 13-22.
- [37] LOH, P. (2015). Statistical Consistency and Asymptotic Normality for High-Dimensional Robust M-Estimators. **arXiv:1501.00312**
- [38] LONG, J. S. and ERVIN, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Model. *The American Statistician* **54**(3), 217-224.

- [39] MACKINNON, J. and WHITE, H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics* **29**, 305–325.
- [40] MAMMEN, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics* **21** (1), 255–285.
- [41] NEWEY, W. K. and WEST, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* **55** (3), 703–708.
- [42] O’HAGAN, A. (2013). Bayesian Inference with Misspecified Models: Inference about what? *Journal of Statistical Planning and Inference* **143**, 1643–1648.
- [43] 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.
- [44] 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>.
- [45] STIGLER, S. M. (2001). Ancillary History. In *State of the Art in Probability and Statistics: Festschrift for W. R. van Zwet* (M. DeGunst, C. Klaassen and A. van der Vaart, eds.), 555–567.
- [46] 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.
- [47] WALKER, S. G. (2013). Bayesian Inference with Misspecified Models. *Journal of Statistical Planning and Inference* **143**, 1621–1633.
- [48] WASSERMAN, L. (2011). Low Assumptions, High Dimensions. *Rationality, Markets and Morals (RMM)* **2** (11), 201–209 (www.rmm-journal.de).
- [49] WEBER, N.C. (1986). The Jackknife and Heteroskedasticity (Consistent Variance Estimation for Regression Models). *Economics Letters* **20**, 161–163.
- [50] WHITE, H. (1980a). Using Least Squares to Approximate Unknown Regression Functions. *International Economic Review* **21** (1), 149–170.
- [51] WHITE, H. (1980b). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **48**, 817–838.
- [52] WHITE, H. (1981). Consequences and Detection of Misspecified Nonlinear Regression Models. *Journal of the American Statistical Association* **76** (374), 419–433.
- [53] WHITE, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**, 1–25.
- [54] WHITE, H. (1994). *Estimation, Inference and Specification Analysis*. Econometric Society Monographs No. 22. Cambridge, GB: Cambridge University Press.
- [55] WU, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics* **14** (4), 1261–1295.

	$\hat{\beta}_j$	SE_{lin}	SE_{boot}	SE_{sand}	$\frac{SE_{boot}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{boot}}$	t_{lin}	t_{boot}	t_{sand}
(Intercept)	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, SE_{boot} and 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, SE_{boot} and SE_{sand} are larger than their linear models cousin 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, SE_{boot} and SE_{sand} are less than 3/4 of SE_{lin} for TAX. For several regressors there is no major discrepancy among all three standard errors: ZN, NOX, B, and even for CRIM, SE_{lin} falls between the slightly discrepant values of SE_{boot} and SE_{sand} .

Table 7 compares standard errors after the

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

Table 8 illustrates the RAV test for the Boston Housing data. Values of \hat{RAV}_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 5 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 β and σ^2 , hence conditioning on \mathbf{X} produces valid frequentist inference for these parameters (Cox and Hinkley 1974, Example 2.27). Indeed, with a suitably general definition of ancillarity, it can be shown that in *any* regression model the regressors form an ancillary. To see this we need an extended definition of ancillarity that includes nuisance parameters. The ingredients and conditions are as follows:

	$\hat{\beta}_j$	SE_{lin}	SE_{boot}	SE_{sand}	$\frac{SE_{boot}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{boot}}$	t_{lin}	t_{boot}	t_{sand}
(Intercept)	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.

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

TABLE 8

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

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

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

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

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

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

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

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

As explained in Section 5, the problem with the ancillarity argument is that it holds only when the regression model is correct. In practice, whether models are correct is never known.

APPENDIX C: ADJUSTMENT

C.1 Adjustment in Populations

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

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

and let

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

The response Y can be adjusted similarly, and we may denote it by $Y_{\bullet-j}$ to indicate that X_j is not among the adjustors, which is implicit in the adjustment of X_j .

C.2 Adjustment in Samples

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

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

and let

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

(Note the use of hat notation “ $\hat{\cdot}$ ” to distinguish it from population-based adjustment “ \bullet ”.) The response vector \mathbf{Y} can be sample-adjusted similarly, and we may denote it by $\mathbf{Y}_{\hat{\cdot}-j}$ to indicate that \mathbf{X}_j is not among the adjustors.

APPENDIX D: PROOFS

D.1 Precise Non-Ancillarity Statements and Proofs for Section 5

Lemma: *The functional $\beta(\mathbf{P})$ depends on \mathbf{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 $\exists \mathbf{P}_1, \mathbf{P}_2 : \beta(\mathbf{P}_1) \neq \beta(\mathbf{P}_2)$ is driven solely by differences in the regressor distributions $\mathbf{P}_1(d\vec{x})$ and $\mathbf{P}_2(d\vec{x})$ because \mathbf{P}_1 and \mathbf{P}_2 share the mean function $\mu_0(\cdot)$ while their conditional noise distributions are irrelevant by the above lemma.

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

$$\mathbf{P}_1(d\vec{x}) = \mathbf{P}_2(d\vec{x}), \quad \mu_1(\vec{X}) \stackrel{\mathbf{P}_{1,2}}{=} \mu_2(\vec{X}) \quad \implies \quad \beta(\mathbf{P}_1) = \beta(\mathbf{P}_2).$$

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

$$\begin{aligned} \mu_0(\cdot) \text{ is nonlinear} &\implies \exists \mathbf{P}_1, \mathbf{P}_2 : \beta(\mathbf{P}_1) \neq \beta(\mathbf{P}_2), \\ \mu_0(\cdot) \text{ is linear} &\implies \forall \mathbf{P}_1, \mathbf{P}_2 : \beta(\mathbf{P}_1) = \beta(\mathbf{P}_2). \end{aligned}$$

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

$$\begin{aligned} \mu_0(\cdot) \text{ is nonlinear} &\implies \exists \mathbf{P}_1, \mathbf{P}_2 : \beta(\mathbf{P}_1) \neq \beta(\mathbf{P}_2), \\ \mu_0(\cdot) \text{ is linear} &\implies \forall \mathbf{P}_1, \mathbf{P}_2 : \beta(\mathbf{P}_1) = \beta(\mathbf{P}_2). \end{aligned}$$

The linear case is trivial: if $\mu_0(\vec{X})$ is linear, that is, $\mu_0(\vec{x}) = \beta'\vec{x}$ for some β , then $\beta(\mathbf{P}) = \beta$ irrespective of $\mathbf{P}(d\vec{x})$. The nonlinear case is proved as follows: For any set of points $\vec{x}_1, \dots, \vec{x}_{p+1} \in \mathbb{R}^{p+1}$ in general position and with 1 in the first coordinate, there exists a unique linear function $\beta'\vec{x}$ through the values of $\mu_0(\vec{x}_i)$. Define $\mathbf{P}(d\vec{x})$ by putting mass $1/(p+1)$ on each point; define the conditional distribution $\mathbf{P}(dy | \vec{x}_i)$ as a point mass at $y = \mu_o(\vec{x}_i)$; this defines \mathbf{P} such that $\beta(\mathbf{P}) = \beta$. Now, if $\mu_0(\cdot)$ is nonlinear, there exist two such sets of points with differing linear functions $\beta_1'\vec{x}$ and $\beta_2'\vec{x}$ to match the values of $\mu_0(\cdot)$ on these two sets; by following the preceding construction we obtain \mathbf{P}_1 and \mathbf{P}_2 such that $\beta(\mathbf{P}_1) = \beta_1 \neq \beta_2 = \beta(\mathbf{P}_2)$.

D.2 RAV Decomposition

Lemma D.2: *RAV Decomposition.*

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

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

D.3 Proof of the RAV-Range Proposition in Section 9.3.5

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

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

Here are some corollaries that follow from the proposition:

- If, for example, $X_{j\bullet} \sim U[-1, +1]$ is uniformly distributed, then $E[X_{j\bullet}^2] = 1/3$. Hence the upper bound on the **RAV** is 3 and, asymptotically, the usual standard error will never be too short by more than a factor $\sqrt{3} \approx 1.732$.
- However, when $E[X_{j\bullet}^2]$ is very small compared to $\mathbf{P}\text{-max } X_{j\bullet}^2$, that is, when $X_{j\bullet}$ is highly concentrated around its mean 0, then this approximates the case of an unbounded support and the worst-case **RAV** can be very large.
- If, on the other hand, $E[X_{j\bullet}^2]$ is very close to $\mathbf{P}\text{-max } X_{j\bullet}^2 = c^2$, then $X_{j\bullet}$ approximates a balanced two-point distribution at $\pm c$, and the sandwich and usual standard errors necessarily agree in the limit.

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

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

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

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

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

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

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

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

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

Then the following holds:

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

Proof of part (a): Preliminary observations:

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

For $p(t) > 0$ we have

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

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

Proof of part (b): Preliminary observations:

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

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

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

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

Lemma D.3.2:

(a) Define a one-parameter family

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

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

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

(b) Define a one-parameter family

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

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

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

Proof of part (a): Preliminary observations:

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

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

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

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

Proof of part (b): Preliminary observations:

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

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

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

D.4 Details for Figure 6

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

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

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

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

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

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

and hence

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

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

D.5 Proof of Asymptotic Normality of \mathbf{RAV}_j , Section 9.4.2

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

$$\{X_{i,j\circ}\}_{i=1\dots N} \xrightarrow{\mathcal{D}} \mathbf{X}_{j\bullet} \stackrel{\mathcal{D}}{=} X_{i,j\bullet} \quad (N \rightarrow \infty).$$

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

$$(44) \quad \mathbf{RAV}_j = \frac{\frac{1}{N} \langle (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}})^2, \mathbf{X}_{j\bullet}^2 \rangle}{\frac{1}{N} \|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}\|^2 \frac{1}{N} \|\mathbf{X}_{j\bullet}\|^2}.$$

For the denominators it is easy to show that

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

For the numerator a CLT holds based on

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

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

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

The second equality on each line holds under the null hypothesis of independent δ and $\vec{\mathbf{X}}$. For the variance one observes that we assume that $\{(Y_i, \vec{\mathbf{X}}_i)\}_{i=1\dots N}$ to be i.i.d. sampled pairs, hence $\{(\delta_i^2, X_{i,j\bullet}^2)\}_{i=1\dots N}$ are N i.i.d. sampled pairs as well. Using the denominator terms (45) and Slutsky's theorem, we arrive at the first version of the CLT for $\mathbf{R}\hat{\mathbf{A}}\mathbf{V}_j$:

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

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

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

Details for the numerator (46), using notation of Sections C.1 and C.2, in particular $\mathbf{X}_{j\bullet} = \mathbf{X}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j\bullet}$ and $\mathbf{X}_{j\hat{\bullet}} = \mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j\hat{\bullet}}$:

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

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

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

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

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

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

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

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

We proceed to the 6 terms in (47) that contain at least two β -subterms or one β -subterm squared. For brevity we treat one term in detail and assume that the reader will be convinced that the other 5 terms can be dealt with similarly. Here is one such term, again scaled for CLT purposes:

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

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

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

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

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

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

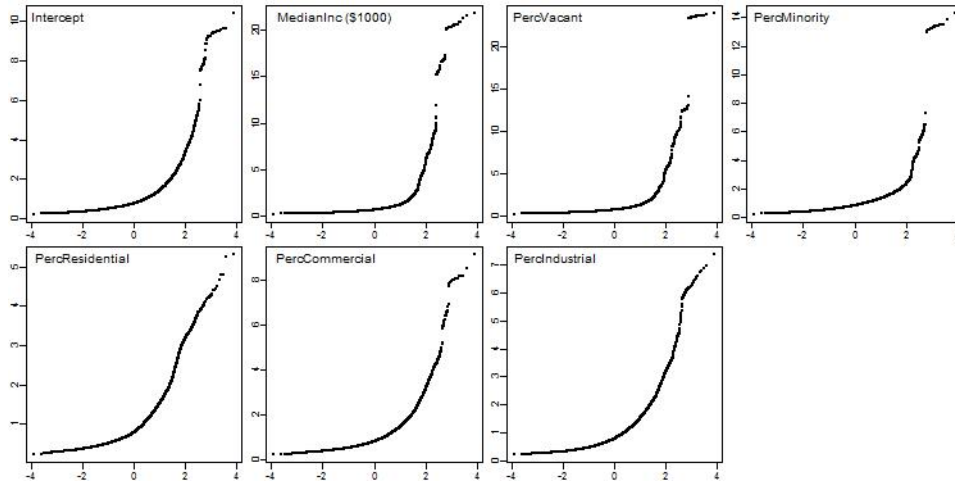


FIG 9. Permutation distributions of $\hat{R}\hat{A}V_j$ for the LA Homeless Data

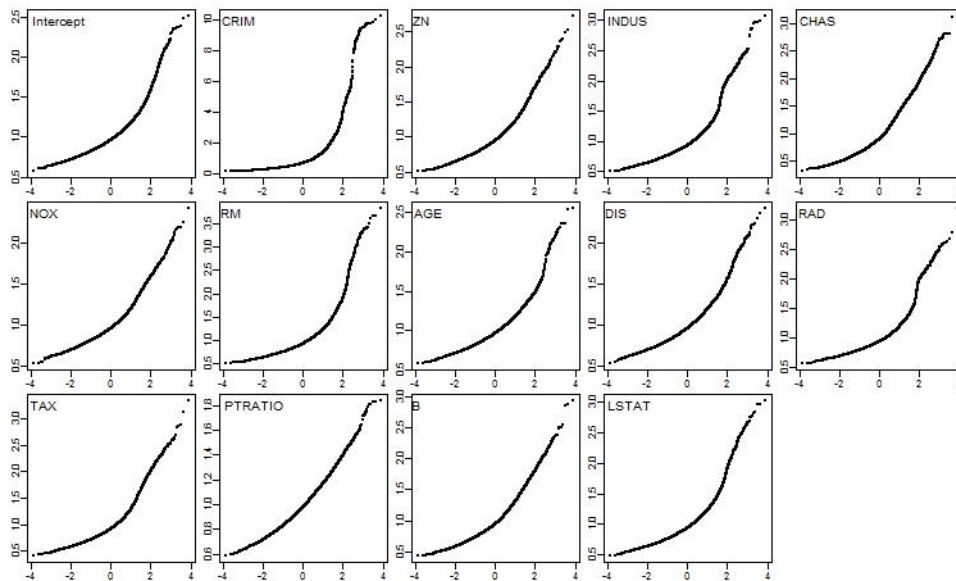


FIG 10. Permutation distributions of $\hat{R}\hat{A}V_j$ for the Boston Housing Data