The Conspiracy of Random Predictors and Model Violations against Classical Inference in Regression

A. Buja, R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, K. Zhang, L. Zhao

March 10, 2014

Dedicated to Halbert White (†2012)

Abstract

We review the early insights of Halbert White who over thirty years ago inaugurated a form of statistical inference for regression models that is asymptotically correct even under "model misspecification." This form of inference, which is pervasive in econometrics, relies on the "sandwich estimator" of standard error. Whereas the classical theory of linear models in statistics assumes models to be correct and predictors to be fixed, White permits models to be "misspecified" and predictors to be random. Careful reading of his theory shows that it is a synergistic effect — a "conspiracy" — of nonlinearity and randomness of the predictors that has the deepest consequences for statistical inference. It will be seen that the synonym "heteroskedasticity-consistent estimator" for the sandwich estimator is misleading because nonlinearity is a more consequential form of model deviation than heteroskedasticity, and both forms are handled asymptotically correctly by the sandwich estimator. The same analysis shows that a valid alternative to the sandwich estimator is given by the "pairs bootstrap" for which we establish a direct connection to the sandwich estimator. We continue with an asymptotic comparison of the sandwich estimator and the standard error estimator from classical linear models theory. The comparison shows that when standard errors from linear models theory deviate from their sandwich analogs, they are usually too liberal, but occasionally they can be too conservative as well. We conclude by answering questions that would occur to statisticians acculturated to the assumption of model correctness and conditionality on the predictors: (1) Why should we be interested in inference for models that are not correct? (2) What are the arguments for conditioning on predictors, and why might they not be valid? In this review we limit ourselves to linear least squares regression as the demonstration object, but the qualitative insights hold for all forms of regression.

Keywords: Ancillarity of predictors; First and second order incorrect models; Model misspecification

1 Introduction

The classical Gaussian linear model reads as follows:

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

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

A second goal of this exposition is to connect the assumption-lean econometric framework to a form of statistical inference in linear models that is known to statisticians but appreciated by few: the "pairs bootstrap." As the name indicates, the pairs bootstrap consists of resampling pairs $(\vec{\boldsymbol{x}}_i, y_i)$, which contrasts with the "residual bootstrap" which resamples residuals r_i. Among the two, the pairs bootstrap is the less promoted even though asymptotic theory exists to justify both types of bootstrap under different assumptions (Freedman 1981, Mammen 1993). It is intuitively clear that the pairs bootstrap can be asymptotically justified in the asumption-lean framework mentioned above, and for this reason it produces standard error estimates that solve the same problem as the sandwich estimators. Indeed, we establish a connection that shows the sandwich estimator to be the asymptotic limit of the m-of-n pairs bootstrap when $m \to \infty$. An advantage of the pairs bootstrap over the sandwich estimator is that higher-order corrections exist for the former when asymptotic normality is an insufficient approximation to the sampling distribution of estimates. In what follows we will use the general term "**assumption-lean estimators of standard error**" to refer to either the sandwich estimators or the pairs bootstrap estimators of standard error.

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

A fourth goal is to move the ***RAV*** from theory to practice by estimating it as a ratio of the two types of standard errors squared. For this estimate we derive an asymptotic null distribution that can be used in a test for the presence of model violations that invalidate the classical standard error. While it is true that this can be called a "misspecification test" in the tradition of econometrics, we prefer to view it as guidance as to the choice of the better standard error. An interesting question raised by such a test concerns the performance of an inferential procedure that chooses the type of standard error depending on the outcome of the test; to this question we do not currently have an answer. (Similar in spirit is Angrist and Pischke's proposal (2009) to choose the larger of the two standard errors; the resulting procedure forfeits the possibility of detecting the classical standard error as too pessimistic.)

A fifth and final goal is to propose answers to questions and objections that would be natural to statisticians who represent the following lines of thinking:

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

A strong proponent of position 1. is the late David Freedman (2006). We will counter-argue as follows:

1.a Models are always approximations, not truths.

1.b If models are approximations, it is still possible to give regression slopes meaningful interpretations.

1.c If models are approximations, it is prudent to use inference that is less dependent on model correctness.

These arguments will be developed in subsequent sections. Position 2. above has proponents to various degrees. More forceful ones hold that conditioning on the predictors is a necessary consequence of the ancillarity principle while others hold that the principle confers license, not a mandate. The ancillarity principle says in simplified terms that valid inference results by conditioning on statistics whose distribution does not involve the parameters of interest. When the predictors in a regression are random, their distribution is ancillary for the regression parameters, hence conditioning on the predictors is necessary or at least permitted. This argument, however, fails when the parametric model is only an approximation and the predictors are random. It will be seen that under these circumstances the population slopes do depend on the predictor distribution which is hence not ancillary. This effect does not exist when the conditional mean of the response is a linear function of the predictors and/or the predictors are truly nonrandom.

[To be updated] This article continuous as follows: Section xyzshows a data example in which the conventional and bootstrap standard errors of some slopes are off by as much as a factor 2. Section xyzintroduces notation for populations and samples, LS approximations, adjustment operations, nonlinearities, and decompositions of responses. Section xyzdescribes the sampling variation caused by random predictors and nonlinearities. Section xyzcompares three types of standard errors and shows that both the (potentially flawed) conventional and the (correct) unconditional standard errors are inflated by nonlinearities, but Section xyzshows that asymptotically there is no limit to which inflation of the unconditional standard error can exceed inflation of the conventional standard error. Appendix 13 gives intuitive meaning to LS slopes in terms of weighted average slopes found in the data.

2 Discrepancies between Standard Errors Illustrated

The table below shows regression results for a dataset in a sample of 505 census tracts in Los Angeles that has been used to examine homelessness in relation to covariates for demographics and building usage (Berk et al. 2008). We do not intend a careful modeling exercise but show the raw results of linear regression to illustrate the degree to which discrepancies can arise among three types of standard errors: SE_{lin} from linear models theory, SE_{boot} from the pairs bootstrap ($N_{boot} = 100,000$) and SE_{sand} from the sandwich estimator (according to McKinnon and White (1985)). Ratios of standard errors are shown in bold font when they indicate a discrepancy exceeding 10%.

	$\hat{\beta}_j$	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

The ratios SE_{sand}/SE_{boot} show that the standard errors from the pairs bootstrap and the sandwich estimator are in rather good agreement. Not so for the standard errors based on linear models theory: we have $SE_{boot}, SE_{sand} > SE_{lin}$ for the predictors `PercVacant`, `PercCommercial` and `PercIndustrial`, and $SE_{boot}, SE_{sand} < SE_{lin}$ for `Intercept`, `MedianInc ($1000)`, `PercResidential`. Only for `PercMinority` is SE_{lin} off by less than 10% from SE_{boot} and SE_{sand}. The discrepancies affect outcomes of some of the t-tests: Under linear models theory the predictors `PercCommercial` and `PercIndustrial` have commanding t-values of 2.700 and 2.818, respectively, which are reduced to unconvincing values below 1.9 and 1.6, respectively, if the pairs bootstrap or the sandwich estimator are used. On the other hand, for `MedianInc ($K)` the t-value -0.977 from linear models theory becomes borderline significant with the bootstrap or sandwich estimator if the plausible one-sided alternative with negative sign is used.

The second illustration of discrepancies between types of standard errors, shown in the table below, is 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 how meaningful the analysis is and simply 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 only a fraction of about 0.73 of SE_{lin} for `TAX`. Also worth stating is that for several predictors there is no substantial discrepancy among all three standard errors, namely `ZN`, `NOX`, `B`, and even for `CRIM`, SE_{lin} falls between the somewhat discrepant values of SE_{boot} and 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)`	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

Important messages are the following: (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 show substantial discrepancies; (3) these discrepancies are specific to predictors. In what follows we describe how the discrepancies arise from nonlinearities in the conditional mean and/or heteroskedasticities in the conditional variance of the response given the predictors. Furthermore, it will turn out that SE_{boot} and SE_{sand} are asymptotically correct while SE_{lin} is not.

3 Populations and Targets of Estimation

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

A population view of regression with random predictors has as its ingredients random variables X_1, ..., X_p and Y, where Y is singled out as the response. At this point the only assumption is that these variables have a joint distribution

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

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

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

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

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

for the joint distribution of $(Y, \vec{\boldsymbol{X}})$, the marginal distribution of $\vec{\boldsymbol{X}}$, and the conditional distribution of Y given $\vec{\boldsymbol{X}}$, respectively. Nonsingularity of the predictor covariance matrix is equivalent to nonsingularity of the cross-moment matrix $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$.

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

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

This is sometimes called the "conditional mean function" or the "response surface". Importantly we do not assume that $\mu(\vec{\boldsymbol{X}})$ is a linear function of $\vec{\boldsymbol{X}}$.

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

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

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

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

The population coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ form a *statistical functional* that is defined for a large class of data distributions $\boldsymbol{P}$. The question of how $\boldsymbol{\beta}(\boldsymbol{P})$ relates to coefficients in the classical linear model (1) will be answered in Section 5.

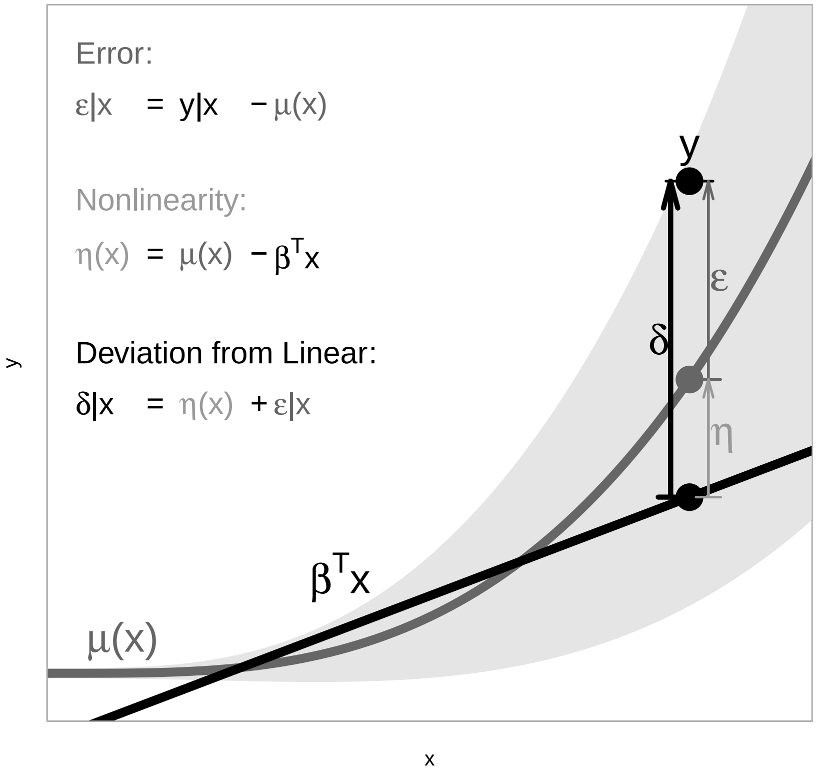

Figure 1: *Illustration of the decomposition* (6).

The population coefficients $\boldsymbol{\beta}(\boldsymbol{P})$ provide also the best linear $L_2(\boldsymbol{P})$ approximation to $\mu(\vec{\boldsymbol{X}})$:

$$\boldsymbol{\beta}(\boldsymbol{P}) \;=\; \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \boldsymbol{E}[(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}^T\vec{\boldsymbol{X}})^2] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}\boldsymbol{E}[\,\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})\,]\,. \tag{5}$$

This fact shows that $\boldsymbol{\beta}(\boldsymbol{P})$ depends on $\boldsymbol{P}$ only in a limited way, as will be spelled out below.

The response Y has the following natural decompositions:

$$\begin{aligned} Y \;&=\; \boldsymbol{\beta}^T\vec{\boldsymbol{X}} \;+\; \underbrace{(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}^T\vec{\boldsymbol{X}})}_{} + \underbrace{(Y - \mu(\vec{\boldsymbol{X}})}_{} \\ &=\; \boldsymbol{\beta}^T\vec{\boldsymbol{X}} \;+\; \underbrace{\eta(\vec{\boldsymbol{X}}) \qquad + \qquad \epsilon}_{} \\ &=\; \boldsymbol{\beta}^T\vec{\boldsymbol{X}} \;+\; \qquad\quad \delta \end{aligned} \tag{6}$$

These equalities define the random variable $\eta = \eta(\vec{\boldsymbol{X}})$, called "nonlinearity", and ϵ, called "error" or "noise", as well $\delta = \epsilon + \eta$, for which there is no standard term so that "linearity deviation" may suffice. Unlike $\eta = \eta(\vec{\boldsymbol{X}})$, the error ϵ and the linearity deviation δ are not functions of $\vec{\boldsymbol{X}}$ alone; if there is a need to refer to the conditional distribution of either given $\vec{\boldsymbol{X}}$, we may write them as $\epsilon|\vec{\boldsymbol{X}}$ and $\delta|\vec{\boldsymbol{X}}$, respectively. An attempt to depict the decompositions (6) for a single predictor is given in Figure 1.

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

$$\boldsymbol{E}[\,\epsilon\,|\vec{\boldsymbol{X}}] \;\overset{\boldsymbol{P}}{=}\; 0, \qquad\qquad \sigma^2(\vec{\boldsymbol{X}}) \;:=\; \boldsymbol{V}[\,\epsilon\,|\vec{\boldsymbol{X}}] \;=\; \boldsymbol{E}[\,\epsilon^2\,|\,\vec{\boldsymbol{X}}] \;\overset{\boldsymbol{P}}{<}\; \infty. \tag{7}$$

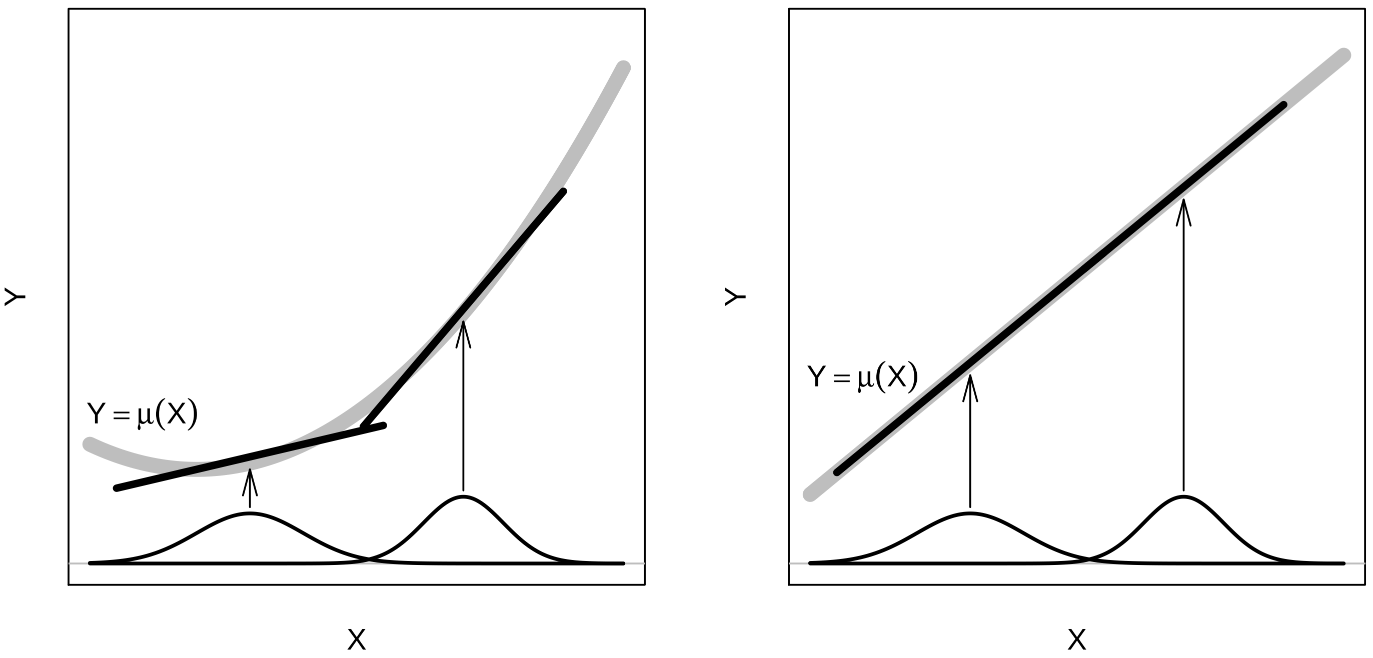

Figure 2: *Illustration of the dependence of the population LS solution on the marginal distribution of the predictors: The left figure shows dependence in the presence of nonlinearity; the right figure shows independence in the presence of linearity.*

We will also need a quantity that describes the total conditional variation of the response around the LS linear function:

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

We refer to it as the "conditional mean squared error" of the population LS function.

Equations (6) above can be given the following *semi-parametric interpretation*:

$$\underbrace{\mu(\vec{\boldsymbol{X}})}_{\textit{semi-parametric part}} \quad = \quad \underbrace{\boldsymbol{\beta}^T\vec{\boldsymbol{X}}}_{\textit{parametric part}} \quad + \quad \underbrace{\eta(\vec{\boldsymbol{X}})}_{\textit{nonparametric part}} \tag{9}$$

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

To make the decomposition (9) identifiable one needs an orthogonality constraint:

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

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

$$\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\eta\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\epsilon\,] \;=\; \boldsymbol{0}, \qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\delta\,] \;=\; \boldsymbol{0}. \tag{10}$$

Proofs: The nonlinearity η is uncorrelated with the predictors because it is the population residual of the regression of $\mu(\vec{\boldsymbol{X}})$ on $\vec{\boldsymbol{X}}$ according to (5). The error ϵ is uncorrelated with $\vec{\boldsymbol{X}}$ because $\boldsymbol{E}[\vec{\boldsymbol{X}}\epsilon] = \boldsymbol{E}[\vec{\boldsymbol{X}}\boldsymbol{E}[\epsilon|\vec{\boldsymbol{X}}]] = \boldsymbol{0}$. Finally, δ is uncorrelated with $\vec{\boldsymbol{X}}$ because $\delta = \eta + \epsilon$.

While the nonlinearity $\eta = \eta(\vec{\boldsymbol{X}})$ is uncorrelated with the predictors, it is not independent from them as it still is a function of them. By comparison, the error ϵ as defined above is *not* independent of the predictors either, but it enjoys a stronger orthogonality property than η: $\boldsymbol{E}[\, g(\vec{\boldsymbol{X}})\, \epsilon\,] = 0$ for all $g(\vec{\boldsymbol{X}}) \in L_2(\boldsymbol{P})$.

Observations:

(0) The error ϵ and the predictors $\vec{\boldsymbol{X}}$ are independent if and only if the conditional distribution of ϵ given $\vec{\boldsymbol{X}}$ is the same across predictor space or, more precisely:

$$\boldsymbol{E}[f(\epsilon)\,|\,\vec{\boldsymbol{X}}] \overset{\boldsymbol{P}}{=} \boldsymbol{E}[f(\epsilon)] \qquad \forall f(\epsilon) \in L_2.$$

(1) The LS functional $\boldsymbol{\beta}(\boldsymbol{P})$ depends on $\boldsymbol{P}$ only through the conditional mean function and the predictor distribution; it does not depend on the conditional error distribution. That is, for two data distributions $\boldsymbol{P}_1(\mathrm{d}y, \mathrm{d}\vec{\boldsymbol{x}})$ and $\boldsymbol{P}_2(\mathrm{d}y, \mathrm{d}\vec{\boldsymbol{x}})$ we have:

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

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

$$\begin{array}{llcl} \text{(2a)} & \mu_0(.) \textit{ is nonlinear} & \Longrightarrow & \exists \boldsymbol{P}_1, \boldsymbol{P}_2 : \;\; \boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2), \\ \text{(2b)} & \mu_0(.) \textit{ is linear} & \Longrightarrow & \forall \boldsymbol{P}_1, \boldsymbol{P}_2 : \;\; \boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}(\boldsymbol{P}_2). \end{array}$$

(For proof details, see Appendix A.1.) In the nonlinear case (2a) the clause $\exists \boldsymbol{P}_1, \boldsymbol{P}_2 : \boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$ is driven solely by differences in the predictor distributions $\boldsymbol{P}_1(\mathrm{d}\vec{\boldsymbol{x}})$ and $\boldsymbol{P}_2(\mathrm{d}\vec{\boldsymbol{x}})$ because $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ share the mean function $\mu_0(.)$ while their conditional error distributions are irrelevant by observation (1).

Observation (2) is much more easily explained with a graphical illustration: Figure 2 shows a single predictor situation with a nonlinear and a linear mean function and two predictor distributions. The two population LS lines for the two predictor distributions differ in the nonlinear case and they are identical in the linear case.

The relevance of observation (2) is that in the presence of nonlinearity the LS functional $\boldsymbol{\beta}(\boldsymbol{P})$ depends on the predictor distribution, hence the predictors are not ancillary for $\boldsymbol{\beta}(\boldsymbol{P})$. The practical implication is this: Consider two empirical studies that use the same predictor and response variables. If their statistical inferences about $\boldsymbol{\beta}(\boldsymbol{P})$ seem superficially contradictory, there may yet be no contradiction if the response surface is nonlinear and the predictor distributions in the two studies differ: it is then possible that the two studies differ in their targets of estimation, $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$. A difference in predictor distributions in two studies can be characterized as a difference in the "range of possibilities." This is illustrated in Figure 3, again with a one-dimensional predictor. Differences in the range of possibilities, however, become the more likely and the less detectable the higher the dimension of predictor space is.

At this point one might argue against a semi-parametric interpretation of linear regression that allows nonlinearities in the response surface. One may hold that nonlinearities render the results from linear regression too dependent on the range of possibilities, and consequently one may be tempted to return to the idea that linear regression should only be applied when it originates from a well-specified linear model. Against this view one may counter-argue that the emergence of nonlinearity may well be a function of the range of possibilities: For studies that sample from a joint distribution $\boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}})$ with large support in predictor space it is more likely that nonlinearity

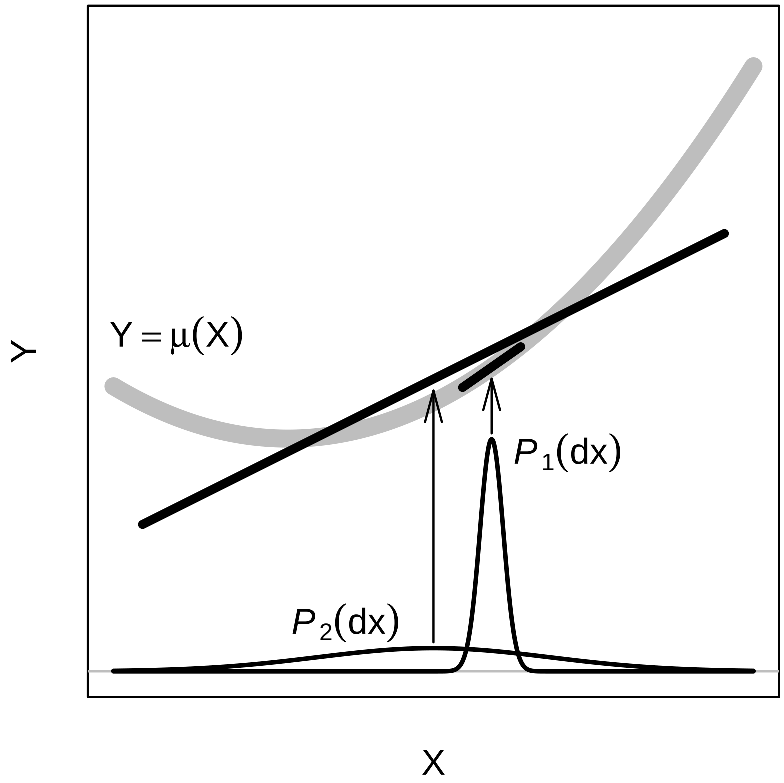

Figure 3: *Illustration of the interplay between the "range of possibilities" and nonlinearity: Over the small support of $\boldsymbol{P}_1$ the nonlinearity will be undetectable and immaterial for realistic sample sizes, whereas over the extended support of $\boldsymbol{P}_2$ the nonlinearity is more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

becomes an important factor because linearity is unlikely to hold globally even if it provides good approximations locally. This is a general fact that cannot be argued away by appeal to "substantive theory" because even the best of theories have limited ranges of validity as has been the experience even with the hardest of theories, those of physics. It therefore seems justified to develop inferential tools for the linear approximation to a nonlinear response surface even if this approximation depends on the range of possibilities. — Figure 3 illustrates the impact of the range of possibilities on $\boldsymbol{\beta}(\boldsymbol{P})$ in the presence of nonlinearity.

4 Observational Datasets and Estimation

The term "observational data" means in this context "cross-sectional data" consisting of i.i.d. cases $(Y_i, X_{i,1}, ..., X_{i,p})$ drawn from a joint multivariate distribution $\boldsymbol{P}(\mathrm{d}y, \mathrm{d}x_1, ..., \mathrm{d}x_p)$ $(i = 1, 2, ..., N)$. We collect the predictors of case i in a column $(p+1)$-vector $\vec{\boldsymbol{X}}_i = (1, X_{i,1}, ..., X_{i,p})^T$, prepended with 1 for an intercept. We stack the N samples to form random column N-vectors and a random predictor $N \times (p+1)$-matrix:

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

Similarly we stack the values $\mu(\vec{\boldsymbol{X}}_i)$, $\eta(\vec{\boldsymbol{X}}_i)$, $\epsilon_i = Y_i - \mu(\vec{\boldsymbol{X}}_i)$, δ_i, and $\sigma(\vec{\boldsymbol{X}}_i)$ to form random column N-vectors:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu(\vec{\boldsymbol{X}}1) \\ .. \\ .. \\ \mu(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta(\vec{\boldsymbol{X}}_1) \\ .. \\ .. \\ \eta(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ .. \\ .. \\ \epsilon_N \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ .. \\ .. \\ \delta_N \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma(\vec{\boldsymbol{X}}_1) \\ .. \\ .. \\ \sigma(\vec{\boldsymbol{X}}_N) \end{bmatrix}. \tag{11}$$

The definitions of $\eta(\vec{\boldsymbol{X}})$, ϵ and δ in (6) translate to vectorized forms:

$$\boldsymbol{\eta} \;=\; \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} \;=\; \boldsymbol{Y} - \boldsymbol{\mu}, \qquad \boldsymbol{\delta} \;=\; \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}. \tag{12}$$

It is important to keep in mind the distinction between population and sample properties. In particular, the N-vectors $\boldsymbol{\delta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are *not* orthogonal to the predictor columns $\boldsymbol{X}_j$ in the sample. Writing $\langle \cdot, \cdot \rangle$ for the usual Euclidean inner product on $\mathbb{R}^N$, we have in general $\langle \boldsymbol{\delta}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\epsilon}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\eta}, \boldsymbol{X}_j \rangle \neq 0$, even though the associated random variables are orthogonal to X_j in the population: $\boldsymbol{E}[\,\delta X_j] = \boldsymbol{E}[\,\epsilon X_j] = \boldsymbol{E}[\,\eta(\vec{\boldsymbol{X}}) X_j] = \;0$.

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

$$\hat{\boldsymbol{\beta}} \;=\; (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^T \;=\; \mathrm{argmin}_{\tilde{\boldsymbol{\beta}}} \, \|\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2 \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,. \tag{13}$$

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

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

The vector $\boldsymbol{r}$ of residuals is of course distinct from the vector $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ as the latter arises from $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$.

5 Decomposition of the LS Estimate According to Two Sources of Variation

When the predictors are random and linear regression is interpreted semi-parametrically as the extraction of the linear part of a nonlinear response surface, the sampling variation of the LS estimate $\hat{\boldsymbol{\beta}}$ can be additively decomposed into two components: one component due to error $\boldsymbol{\epsilon}$ and another component due to nonlinearity interacting with randomness of the predictors. This decomposition is a direct reflection of the decomposition $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$, according to (6) and (12). We give elementary asymptotic normality statements for each part of the decomposition. The relevance of the decomposition is that it explains what the pairs bootstrap estimates, while the associated asymptotic normalities are necessary to justify the pairs bootstrap.

In the classical linear models theory, which is conditional on $\boldsymbol{X}$, the target of estimation is $\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$. When $\boldsymbol{X}$ is treated as random and nonlinearity is permitted, the target of estimation is the population LS solution $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ defined in (4). In this case, $\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$ is a random vector that sits between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; (\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \;+\; (\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta}) \tag{14}$$

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

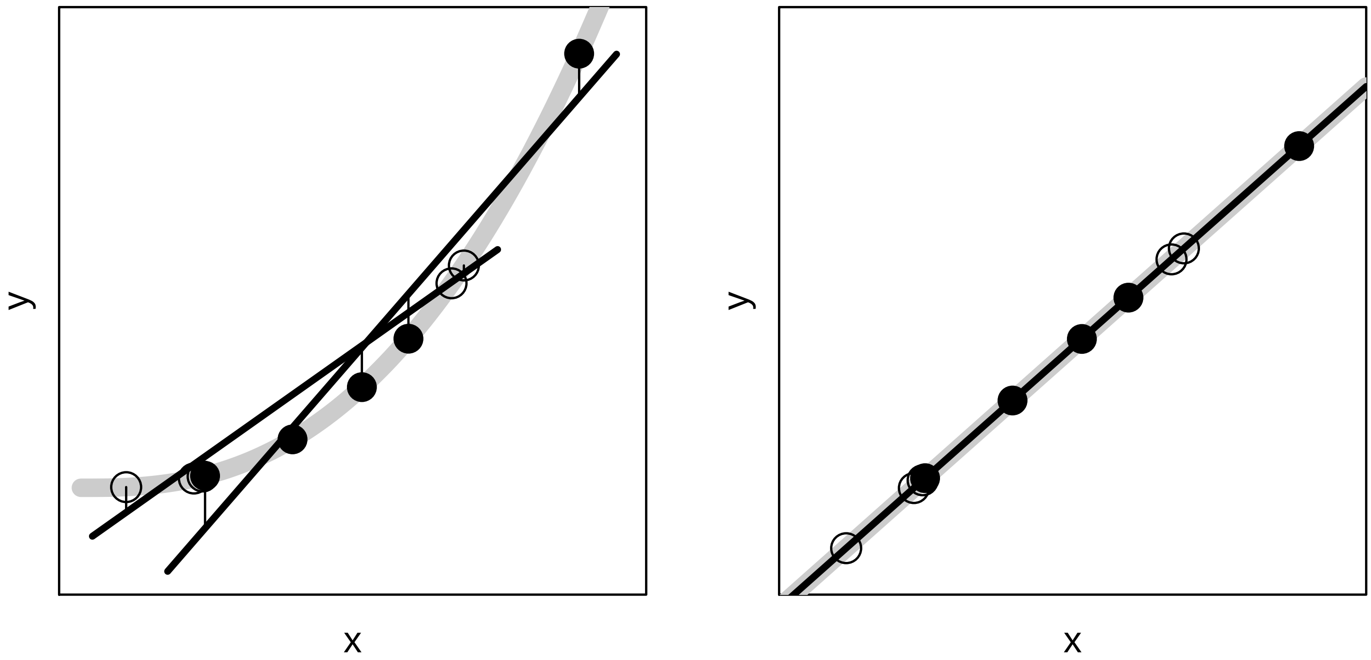

Figure 4: *Error-less Response: The filled and the open circles represent two "datasets" from the same population. The x-values are random; the y-values are a deterministic function of x: $y = \mu(x)$ (shown in gray).*
Left: The true response $\mu(x)$ is nonlinear; the open and the filled circles have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

Definition and Lemma: *The following quantities will be called "Estimation Offsets" or "EO" for short, and they will be prefixed as follows:*

$$\begin{array}{lrcl} \textit{Total EO}: & \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} & = & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\delta}, \\ \textit{Error EO}: & \hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] & = & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}, \\ \textit{Nonlinearity EO}: & \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta} & = & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}. \end{array} \tag{15}$$

This follows immediately from the decompositions (12), $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X\beta}$, $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X\beta}$, and these facts:

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

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

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

$$\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \boldsymbol{E}[\,\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]] \;+\; \boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]]. \tag{16}$$

This decomposition reflects the estimation decomposition (14) and $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$ in view of (15):

$$\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\delta}\,], \tag{17}$$

$$\boldsymbol{E}[\,\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]] \;=\; \boldsymbol{E}[\,\boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\,|\boldsymbol{X}]\,], \tag{18}$$

$$\boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]] \;=\; \boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\,]. \tag{19}$$

(Note that in general $\boldsymbol{E}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\,] \neq \boldsymbol{0}$ even though $\boldsymbol{E}[\,\boldsymbol{X}^T\boldsymbol{\eta}\,] = \boldsymbol{0}$ and hence $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta} \longrightarrow \boldsymbol{0}$ a.s.)

6 Random $\boldsymbol{X}$ Combined with Nonlinearity as a Source of Sampling Variation

Mainstream statistics is largely about sampling variability due to error (18). The fact that there exists another source of sampling variability is little known: nonlinearity (19) in the presence of random predictors. It may be useful to isolate this source and illustrate it in a situation that is free of error: Consider a response that is a deterministic noiseless but nonlinear function of the predictors: $Y = \eta(\vec{\boldsymbol{X}})$. (This may actually be a realistic situation when outcomes from expensive deterministic simulation experiments are modeled based on inputs.) Assume therefore $\boldsymbol{\epsilon} = \mathbf{0}$ but $\boldsymbol{\eta} \neq \mathbf{0}$, hence there exists sampling variability in $\hat{\boldsymbol{\beta}}$ which is solely due to the nonlinearity $\boldsymbol{\eta}$: $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}$ in conjunction with the randomness of the predictors — the "conspiracy" in the title of this article. Figure 4 illustrates the situation with a single-predictor example by showing the LS lines fitted to two "datasets" consisting of $N = 5$ predictor values each. The random differences between datasets cause the fitted line to exhibit sampling variability under nonlinearity (left hand figure), which is absent under linearity (right hand figure). Compare this figure with the earlier Figure 2: mathematically the effects illustrated in both are identical; Figure 2 shows the effect for different populations (theoretical $\boldsymbol{X}$ distributions) while Figure 4 shows it for different datasets (empirical $\boldsymbol{X}$ distributions). Thus nonlinearity creates complications on two interconnected levels: (1) in the definition of the population LS parameter, which becomes dependent on the predictor distribution, and (2) through the creation of sampling variability due to $\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}]$ which becomes a true random vector.

The case of error free but nonlinear data is of interest to make another point regarding statistical inference: If classical linear models theory conditions on the predictors and assumes erroneously that the response surface is linear, it is not so that the resulting procedures do "not see" see the sampling variability caused by nonlinearity, but they misinterpret it as due to error. The consequences of the confusion of errors and nonlinearities for statistical inference will be examined in Section 10.2. This misinterpretation also seeps into the residual bootstrap as it assumes the residuals to originate from exchangeable errors only. By comparison, the pairs bootstrap gets statistical inference right even in the error-free nonlinear case, at least asymptotically. It receives its justification from central limit theorems.

7 Assumption-Lean Central Limit Theorems

The three EOs arise from the decomposition $\delta = \epsilon + \eta$ (6). The respective CLTs draw on the analogous conditional second moment decomposition $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ (8). The asymptotic variance/covariance matrices have the well-known sandwich form:

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

$$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\Big(\mathbf{0},\ \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\Big) \tag{20}$$

$$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\Big(\mathbf{0},\ \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\Big) \tag{21}$$

$$N^{1/2}\,(\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta}) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\Big(\mathbf{0},\ \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\eta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\Big) \tag{22}$$

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

$$\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] \;=\; \boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T], \qquad \boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] \;=\; \boldsymbol{E}[\,\epsilon^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T], \tag{23}$$

which follows from $m^2(\vec{\boldsymbol{X}}) = \boldsymbol{E}[\,\delta^2|\vec{\boldsymbol{X}}]$ and $\sigma^2(\vec{\boldsymbol{X}}) = \boldsymbol{E}[\,\epsilon^2|\vec{\boldsymbol{X}}]$ according to (7) and (8).

Proof Outline: The three cases follow the same way; we consider the first. Using $\boldsymbol{E}[\,\delta\vec{\boldsymbol{X}}\,] = \boldsymbol{0}$ from (10) we have:

$$\begin{aligned} N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(\tfrac{1}{N}\,\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\tfrac{1}{N^{1/2}}\,\boldsymbol{X}^T\boldsymbol{\delta}\right) \\ &= \left(\tfrac{1}{N}\sum\vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i^T\right)^{-1}\left(\tfrac{1}{N^{1/2}}\sum\vec{\boldsymbol{X}}_i\,\delta_i\right) \\ &\xrightarrow{\mathcal{D}} \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\right) \\ &= \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\delta^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\right)\,, \end{aligned} \tag{24}$$

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

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

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

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

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

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

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

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

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

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

8 The Sandwich Estimator and the M-of-N Pairs Bootstrap

Empirically one observes that standard error estimates obtained from the pairs bootstrap and from the sandwich estimator are generally close to each other. This is intuitively unsurprising as they both estimate the same asymptotic variances. A closer connection between them will be established below.

8.1 The Plug-In Sandwich Estimator of Asymptotic Variance

The simplest form of the sandwich estimator of asymptotic variance is the plug-in version of the asymptotic variance as it appears in the CLT of (20), replacing the hard-to-estimate quantity $m^2(\vec{\boldsymbol{X}})$ with the easy-to-estimate quantity $\delta^2 = (Y - \boldsymbol{\beta}\vec{\boldsymbol{X}})^2$ according to (20). For plug-in one estimates the population expectations $\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$ and $\boldsymbol{E}[\,(Y - \vec{\boldsymbol{X}}^T\boldsymbol{\beta})\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$ with sample means and the population parameter $\boldsymbol{\beta}$ with the LS estimate $\hat{\boldsymbol{\beta}}$. For this we use the notation $\hat{\boldsymbol{E}}[...]$ to express sample means:

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

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

$$\hat{\boldsymbol{AV}}_{sand} \;:=\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\;\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1} \tag{25}$$

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

$$\hat{SE}_{sand}(\hat{\beta}_j) \;:=\; \frac{1}{N^{1/2}}(\hat{\boldsymbol{AV}}_{sand})^{1/2}_{jj}. \tag{26}$$

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

To connect the sandwich estimator (25) to its bootstrap counterpart we need the M-of-N bootstrap whereby the *resample size* M is allowed to differ from the sample size N. It is at this point important not to confuse

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

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

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

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

$$M^{1/2}\,(\boldsymbol{\beta}^*_M - \hat{\boldsymbol{\beta}}) \;\; \xrightarrow{\mathcal{D}} \;\; \mathcal{N}\Big(\mathbf{0},\; \hat{\boldsymbol{E}}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\;\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\Big) \quad (M \to \infty). \tag{27}$$

This is a straight application of the CLT of the previous section to the empirical distribution rather than the actual distribution of the data, where the middle part (the "meat") of the asymptotic formula is based on the empirical counterpart $r_i^2 = (Y_i - \vec{\boldsymbol{X}}_i^T\hat{\boldsymbol{\beta}})^2$ of $\delta^2 = (Y - \vec{\boldsymbol{X}}^T\boldsymbol{\beta})^2$. A comparison of (25) and (27) results in the following:

Observation: *The sandwich estimator* (25) *is the asymptotic variance estimated by the limit of the M-of-N pairs bootstrap as $M \to \infty$ for a fixed sample of size N.*

9 Adjusted Predictors

The adjustment formulas of this section serve to express the slopes of multiple regressions as slopes in simple regressions using adjusted single predictors. The goal is to analyze the discrepancies between the proper and improper standard errors of regression estimates in subsequent sections.

9.1 Adjustment formulas for the population

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

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

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

$$X_{j\bullet} \;=\; X_j - \boldsymbol{\beta}_{j\bullet}^T \vec{\boldsymbol{X}}_{-j}\,. \tag{28}$$

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

$$\beta_j \;=\; \frac{\boldsymbol{E}[\,Y X_{j\bullet}]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]} \;=\; \frac{\boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}})X_{j\bullet}]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]}\,. \tag{29}$$

9.2 Adjustment formulas for samples

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

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

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

$$\boldsymbol{X}_{j\hat{\bullet}} \;=\; \boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{j\hat{\bullet}} \;=\; (\boldsymbol{I} - \boldsymbol{H}_{-j})\boldsymbol{X}_j\,. \tag{30}$$

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

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

With the above notation we can make the following distinctions: $X_{i,j\bullet}$ refers to the i'th i.i.d. replication of the population-adjusted random variable $X_{j\bullet}$, whereas $X_{i,j\hat{\bullet}}$ refers to the i'th component of the sample-adjusted random column $\boldsymbol{X}_{j\hat{\bullet}}$. Note that the former, $X_{i,j\bullet}$, are i.i.d. for $i = 1, ..., N$, whereas the latter, $X_{i,j\hat{\bullet}}$, are not because sample adjustment introduces dependencies throughout the components of the random N-vector $\boldsymbol{X}_{j\hat{\bullet}}$. As $N \to \infty$ for fixed p, however, this dependency disappears asymptotically, and we have for the empirical distribution of the values $\{X_{i,j\hat{\bullet}}\}_{i=1...N}$ the obvious convergence in distribution:

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

Actually, under suitable moment assumptions this holds even in Mallows metrics, meaning that not only empirical frequencies converge to their limiting probabilities but also empirical moments converge to their population moments.

9.3 Adjustment Formulas for Decompositions and Their CLTs

The vectorized formulas for estimation offsets (14) have the following component analogs:

$$
\begin{array}{lrcl}
\textit{Total EO}: & \hat{\beta}_j - \beta_j & = & \dfrac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\delta} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}, \\[2ex]
\textit{Error EO}: & \hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] & = & \dfrac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\epsilon} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}, \\[2ex]
\textit{Nonlinearity EO}: & \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j & = & \dfrac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\eta} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}.
\end{array}
\tag{32}
$$

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

Corollary:

$$
\begin{array}{rcll}
N^{1/2}(\hat{\beta}_j - \beta_j) & \xrightarrow{\mathcal{D}} & \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}}) X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) & = \ \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\,\delta^2 X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) \\[3ex]
N^{1/2}(\hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}]) & \xrightarrow{\mathcal{D}} & \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}}) X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) & \\[3ex]
N^{1/2}(\boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j) & \xrightarrow{\mathcal{D}} & \mathcal{N}\left(0, \dfrac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}}) X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) &
\end{array}
\tag{33}
$$

10 Asymptotic Variances — Proper and Improper

10.1 Proper and Improper Asymptotic Variances Expressed with Adjusted Predictors

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

Here is generic notation that can be used to describe the proper asymptotic variance of $\hat{\beta}_j$ as well as its decomposition into components due to error and due to nonlinearity:

Definition:

$$
\boldsymbol{AV}^{(j)}_{lean}(f^2(\vec{\boldsymbol{X}})) \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}}) X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}
\tag{34}
$$

The proper asymptotic variance of $\hat{\beta}_j$ and its decomposition is therefore according to (33)

$$
\begin{array}{ccccc}
\boldsymbol{AV}^{(j)}_{lean}(m^2(\vec{\boldsymbol{X}})) & = & \boldsymbol{AV}^{(j)}_{lean}(\sigma^2(\vec{\boldsymbol{X}})) & + & \boldsymbol{AV}^{(j)}_{lean}(\eta^2(\vec{\boldsymbol{X}})) \\[3ex]
\dfrac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}}) X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2} & = & \dfrac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}}) X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2} & + & \dfrac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}}) X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}
\end{array}
\tag{35}
$$

The next step is to derive an asymptotic form for the conventional standard error estimate in the assumption-lean framework. This asymptotic form will have the appearance of an asymptotic variance but it is valid only in the assumption-loaded framework of first and second order well-specification. This "improper" standard error depends on an estimate $\hat{\sigma}^2$ of the error variance, usually $\hat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(N{-}p{-}1)$. In an assumption-lean context, with both heteroskedastic error variance and nonlinearity, $\hat{\sigma}^2$ has the following limit:

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

Standard error estimates are therefore given by

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \hat{\sigma}^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \qquad \hat{SE}^2_{lin}[\,\hat{\beta}_j] \;=\; \frac{\hat{\sigma}^2}{\|\boldsymbol{X}_{j\bullet}\|^2}. \tag{36}$$

Their scaled limits are (a.s.) under usual assumptions as follows:

$$N\,\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \;\overset{N\to\infty}{\longrightarrow}\; \boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,]\;\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}, \qquad N\,\hat{SE}^2_{lin}[\,\hat{\beta}_j] \;\overset{N\to\infty}{\longrightarrow}\; \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,]}{\boldsymbol{E}[\,X^2_{j\bullet}\,]}. \tag{37}$$

These are the asymptotic expressions that describe the limiting behavior of linear models standard errors in an assumption-lean context. Even though they are not proper asymptotic variances except in an assumption-loaded context, they are intended and used as such. We introduce the following generic notation for improper asymptotic variance where $f^2(\vec{\boldsymbol{X}})$ is again a placeholder for any one among $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$:

Definition:

$$\boldsymbol{AV}^{(j)}_{lin}(f^2(\vec{\boldsymbol{X}})) \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \tag{38}$$

Here is the improper asymptotic variance of $\hat{\beta}_j$ and its decomposition into components due to error and nonlinearity:

$$\begin{aligned} \boldsymbol{AV}^{(j)}_{lin}(m^2(\vec{\boldsymbol{X}})) \;&=\; \boldsymbol{AV}^{(j)}_{lin}(\sigma^2(\vec{\boldsymbol{X}})) \;+\; \boldsymbol{AV}^{(j)}_{lin}(\eta^2(\vec{\boldsymbol{X}})) \\ \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \;&=\; \frac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \;+\; \frac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \end{aligned} \tag{39}$$

We examine next the discrepancies between proper and improper asymptotic variances.

10.2 Comparison of Proper and Improper Asymptotic Variances

It will be shown that the conventional asymptotic variances can be too small or too large to unlimited degrees compared to the proper marginal asymptotic variances. A comparison of asymptotic variances can be done separately for $\sigma^2(\vec{\boldsymbol{X}})$, $\eta^2(\vec{\boldsymbol{X}})$ and $m^2(\vec{\boldsymbol{X}})$. To this end we form the ratios $\boldsymbol{RAV}_j(...)$ as follows:

Definition and Lemma: *Ratios of Proper and Improper Asymptotic Variances*

$$\begin{aligned} \boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) \;&:=\; \frac{\boldsymbol{AV}^{(j)}_{lean}(m^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{lin}(m^2(\vec{\boldsymbol{X}}))} \;=\; \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_{j\bullet}{}^2]} \\ \boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}})) \;&:=\; \frac{\boldsymbol{AV}^{(j)}_{lean}(\sigma^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{lin}(\sigma^2(\vec{\boldsymbol{X}}))} \;=\; \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_{j\bullet}{}^2]} \\ \boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}})) \;&:=\; \frac{\boldsymbol{AV}^{(j)}_{lean}(\eta^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{lin}(\eta^2(\vec{\boldsymbol{X}}))} \;=\; \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_{j\bullet}{}^2]} \end{aligned} \tag{40}$$

The second equality on each line follows from (39) and (35). The ratios in (40) express by how much the improper conventional asymptotic variances need to multiplied to match the proper asymptotic variances. Among the three ratios the relevant one for the overall comparison of improper conventional and proper inference is $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$. For example, if $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) = 4$, say, then, for large sample sizes, the correct marginal standard error of $\hat{\beta}_j$ is about twice as large as the incorrect conventional standard error. In general $\boldsymbol{RAV}_j$ expresses the following:

- If $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) = 1$, the conventional standard error for $\hat{\beta}_j$ is asymptotically correct;
- if $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) > 1$, the conventional standard error for β_j is asymptotically too small/optimistic;
- if $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) < 1$, the conventional standard error for β_j is asymptotically too large/pessimistic.

The ratios $\boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}}))$ and $\boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}}))$ express the degrees to which heteroskedasticity and/or nonlinearity contribute asymptotically to the defects of conventional standard errors. Note, however, that if $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) = 1$ and if hence the conventional standard error is asymptotically correct, it does not imply that the model is well-specified in any sense. For example, heteroskedasticity and nonlinearity can in principle conspire to make $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}}) = m_0^2$ constant while neither $\sigma^2(\vec{\boldsymbol{X}})$ nor $\eta^2(\vec{\boldsymbol{X}})$ are constant. In this case the conventional standard error is asymptotically correct yet the model is "misspecified." Well-specification with $\sigma^2(\boldsymbol{X}) = \sigma_0^2$ constant and $\eta = 0$ is a sufficient but not necessary condition for asymptotic validity of the conventional standard error.

10.3 Meaning and Range of the *RAV*

The three $\boldsymbol{RAV}$s are inner products between the normalized squared quantities

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

on the one hand, and the normalized squared adjusted predictor

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

on the other hand. These inner products, however, are *not* correlations, and they are *not* bounded by +1; their natural bounds are rather 0 and ∞, and both can generally be approached to any degree when allowing $\sigma^2(\vec{\boldsymbol{X}})$, $\eta(\vec{\boldsymbol{X}})$, and hence $m^2(\vec{\boldsymbol{X}})$ to vary. We continue with generic notation $f^2(\vec{\boldsymbol{X}})$ to mean either of $\sigma^2(\vec{\boldsymbol{X}})$, $\eta^2(\vec{\boldsymbol{X}})$ and $m^2(\vec{\boldsymbol{X}})$:

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

$$\sup_f \boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) = \infty\,. \tag{41}$$

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

$$\inf_f \boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) = 0\,. \tag{42}$$

For a proof see Appendix A.6. To reach the limits 0 and $+\infty$ it is sufficient to consider sequences of functions of $X_{j\bullet}$ only: $f(\vec{\boldsymbol{X}}) = f(X_{j\bullet})$. This is obvious from the following fact which shows that one can further reduce the problem to functions of ${X_{j\bullet}}^2$, which, however, we continue to write as $f(X_{j\bullet})$:

$$\boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) \;=\; \frac{\boldsymbol{E}[f_j^2(X_{j\bullet})\, {X_{j\bullet}}^2]}{\boldsymbol{E}[f_j^2(X_{j\bullet}]\; \boldsymbol{E}[{X_{j\bullet}}^2]} \qquad \text{for} \qquad f_j^2(X_{j\bullet}) \;=\; \boldsymbol{E}[f^2(\vec{\boldsymbol{X}}) \,|\, {X_{j\bullet}}^2]. \tag{43}$$

The problem then boils down to a single predictor situation in $X = X_{j\bullet}$ which conveniently lends itself to graphical illustration: Figure 5 shows a family of functions $f^2(x)$ that interpolates the range of $\boldsymbol{RAV}$ values from 0 to ∞.

Even though the $\boldsymbol{RAV}$ is not a correlation, it is nevertheless a measure of association between $f^2(\vec{\boldsymbol{X}})$ and ${X_{j\bullet}}^2$. It exists for constant but nonzero $f(\vec{\boldsymbol{X}})^2$, in which case $\boldsymbol{RAV} = 1$. It indicates a positive association between $f^2(\vec{\boldsymbol{X}})$ and ${X_{j\bullet}}^2$ for $\boldsymbol{RAV} > 1$ and a negative association for $\boldsymbol{RAV} < 1$. This is borne out by the figures: large values $\boldsymbol{RAV} > 1$ are obtained when $f^2(X_{j\bullet})$ is large for $X_{j\bullet}$ far from zero, and small values $\boldsymbol{RAV} < 1$ are obtained when $f^2(X_{j\bullet})$ is large for $X_{j\bullet}$ near zero.

These examples allow us to train our intuitions about the types of heteroskedasticities and nonlinearities that drive the $\boldsymbol{RAV}$:

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

These facts are illustrated in Figures 6 and 7. To the authors these facts also suggest the following: large values $\boldsymbol{RAV}_j > 1$ are generally more likely than small values $\boldsymbol{RAV}_j < 1$ because both large conditional variances and nonlinearities are often more pronounced in the extremes of predictor distributions. This seems particularly natural for nonlinearities which in the simplest cases will be convex or concave. None of this is more than practical common sense, however, and the occasional exception exists as indicated by the data example of Section 2.

10.4 Combining the $\boldsymbol{RAV}$s from Heteroskedasticity and Nonlinearity

In the end it is $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ that determines the discrepancy between inferences from assumption-loaded linear models theory and assumption-lean approaches based on pairs bootstrap or sandwich estimators. The following observation can be used to examine the behavior of $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$ based on its components $\boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}}))$ and $\boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}}))$:

Observation: *Define weights*

$$w_\sigma = \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}\,, \qquad w_\eta = \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]}\,, \tag{44}$$

so that $w_\sigma + w_\eta = 1$. Then

$$\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) \;=\; w_\sigma\, \boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}})) + w_\eta\, \boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}}))\,. \tag{45}$$

Corollary: *If* $m_t^2(\vec{\boldsymbol{X}}) = \sigma_t^2(\vec{\boldsymbol{X}}) + \eta_t^2(\vec{\boldsymbol{X}})$ *is a one-parameter family of scenarios, and if* $w_{\sigma_t} = w_\sigma > 0$ *and* $w_{\eta_t} = w_\eta > 0$ *from* (44) *are fixed and strictly positive, then we have*

$$\lim_t \boldsymbol{RAV}_j(m_t^2(\vec{\boldsymbol{X}})) = \infty \quad \textit{if} \quad \lim_t \boldsymbol{RAV}_j(\sigma_t^2(\vec{\boldsymbol{X}})) = \infty \quad \textbf{\textit{or}} \quad \lim_t \boldsymbol{RAV}_j(\eta_t^2(\vec{\boldsymbol{X}})) = \infty;$$

$$\lim_t \boldsymbol{RAV}_j(m_t^2(\vec{\boldsymbol{X}})) = 0 \quad \textit{if} \lim_t \boldsymbol{RAV}_j(\sigma_t^2(\vec{\boldsymbol{X}})) = 0 \quad \textbf{\textit{and}} \quad \lim_t \boldsymbol{RAV}_j(\eta_t^2(\vec{\boldsymbol{X}})) = 0.$$

Thus heteroscedasticity $\sigma^2(\vec{\boldsymbol{X}})$ or nonlinearity $\eta^2(\vec{\boldsymbol{X}})$ can each individually cause $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$ to explode, and each can provide a floor that prevents $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$ from dropping to 0. The corollary therefore adds evidence that large values $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) \gg 1$ seem more easily possible than small values $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) \ll 1$, hence conventional linear models inferences seem more likely to be too optimistic/liberal than too pessimistic/conservative.

11 The Sandwich Estimator in Adjusted Form

The adjustment versions (33) of the CLTs can be used to rewrite the sandwich estimator of asymptotic variance by replacing expectations $\boldsymbol{E}[...]$ with means $\hat{\boldsymbol{E}}[...]$, the population parameter $\boldsymbol{\beta}$ with its estimate $\hat{\boldsymbol{\beta}}$, and population adjustment $X_{j\bullet}$ with sample adjustment $\boldsymbol{X}_{j\hat{\bullet}}$:

$$\hat{\boldsymbol{AV}}_{sand}^{(j)} \;=\; \frac{\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2 X_{j\hat{\bullet}}{}^2]}{\hat{\boldsymbol{E}}[\,X_{j\hat{\bullet}}{}^2]^2} \;=\; N\,\frac{\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^4} \tag{46}$$

The squaring of N-vectors is meant to be coordinate-wise. The formula (46) is not a new estimator of asymptotic variance; rather, it is an algebraically equivalent re-expression of the diagonal elements of $\hat{\boldsymbol{AV}}_{sand}$ in (25) above: $\hat{\boldsymbol{AV}}_{sand}^{(j)} = (\hat{\boldsymbol{AV}}_{sand})_{j,j}$. The sandwich standard error estimate (26) can therefore be written as follows:

$$\hat{SE}_{sand}(\hat{\beta}_j) \;=\; \frac{\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle^{1/2}}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}. \tag{47}$$

For comparison the usual standard error estimate from linear models theory is (36):

$$\hat{SE}_{lin}(\hat{\beta}_j) \;=\; \frac{\hat{\sigma}}{\|\boldsymbol{X}_{j\hat{\bullet}}\|}, \tag{48}$$

where as usual $\hat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(N–p–1)$.

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

$$\hat{\boldsymbol{RAV}}_j \;:=\; N\,\frac{\langle (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2, \boldsymbol{X}_{j\hat{\bullet}}{}^2\rangle}{\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\,\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,. \tag{49}$$

This diagnostic quantity can be leveraged as a test statistic because under the assumption of linearity and homoskedasticity a centered and scaled version has an asymptotic null distribution as follows:

$$N^{1/2}\,(\hat{\boldsymbol{RAV}}_j - 1) \;\;\xrightarrow{\mathcal{D}}\;\; \mathcal{N}(0,1) \qquad (N \to \infty)\,. \tag{50}$$

The resulting test could be interpreted as a "per predictor misspecification test," but more pragmatically it could provide guidance as to which standard error estimate to use: (47) in case of rejection and (48) in case of retention of the null hypothesis of linearity and homoskedasticity. This has been suggested by White (1980??) for another form of a misspecification test, but the statistical performance of such pre-test procedures is not known.

12 Justification of Conditioning on the Predictors — and its Failure

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

(1) $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$: the parameters, where $\boldsymbol{\psi}$ is of interest and $\boldsymbol{\lambda}$ is nuisance;

(2) $\boldsymbol{S} = (\boldsymbol{T}, \boldsymbol{A})$: a sufficient statistic with values $(\boldsymbol{t}, \boldsymbol{a})$;

(3) $p(\boldsymbol{t}, \boldsymbol{a};\, \boldsymbol{\psi}, \boldsymbol{\lambda}) = p(\boldsymbol{t}\,|\,\boldsymbol{a};\, \boldsymbol{\psi})\; p(\boldsymbol{a};\, \boldsymbol{\lambda})$: the condition that makes $\boldsymbol{A}$ an ancillary.

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

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

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

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

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

where $\boldsymbol{X}$ acts as the ancillary $\boldsymbol{A}$ and $p_{\boldsymbol{X}}$ as the mixing distribution $p(\boldsymbol{a}\,|\boldsymbol{\lambda})$ with a "nonparametric" nuisance parameter that allows largely arbitrary distributions for the predictors. (The predictor distribution should grant identifiability of $\boldsymbol{\theta}$ in general, and non-collinearity in linear models in particular.) The literature does not seem to be rich in crisp definitions of ancillarity, but see, for example, Cox and Hinkley (1974, p. 32-33). For the interesting history of ancillarity see the articles by Stigler (1986) and Aldrich (2005).

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, yet statisticians fit them just the same and interpret them with statistical inference that assumes the truth of the models. As we have seen such inference can be faulty due to the presence of nonlinearities and heteroskedasticities. The most egregious violation of the ancillarity concept is caused by nonlinearity ("first order misspecification") which requires the estimated parameter to be interpreted as a statistical functional $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ that depends on the predictor distribution (Section 3). Nonlinearity therefore creates a link between parameters $\boldsymbol{\beta}(\boldsymbol{P})$ and predictor distributions $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, violating the postulate of ancillarity in a surprising direction: It is *not* the case that the predictor distribution is dependent on the parameter $\boldsymbol{\beta}$, rather, the parameter $\boldsymbol{\beta}$ is a function of the predictor distribution, $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$, as illustrated in Figure 2.

13 The Meaning of Regression Slopes in the Presence of Nonlinearity

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

- **Sample estimates**: In a multiple regression based on a sample of size N, consider the LS estimate $\hat{\beta}_j$: this is the empirical simple regression slope through the origin with regard to the empirically adjusted predictor $\boldsymbol{X}_{j\hat{\bullet}}$ (for $j \neq 0$ as we only consider actual slopes, not the intercept, but assume the presence of an intercept). We write $(x_1, ..., x_N)^T$ for $\boldsymbol{X}_{j\hat{\bullet}}$, as well as $(y_1, ..., y_N)^T$ for the response vector $\boldsymbol{Y}$ and $\hat{\beta}$ for the LS estimate $\hat{\beta}_j$. Then the representation of $\hat{\beta}$ as a weighted average of case-wise slopes is

  $$\hat{\beta} \;=\; \sum_i w_i \, b_i \,, \quad \text{where} \quad b_i \;:=\; \frac{y_i}{x_i} \quad \text{and} \quad w_i \;:=\; \frac{x_i^2}{\sum_{i'} x_{i'}^2} \tag{51}$$

  are case-wise slopes and weights, respectively.

  The representation of $\hat{\beta}$ as a weighted average of pairwise slopes is

  $$\hat{\beta} \;=\; \sum_{ik} w_{ik} \, b_{ik} \,, \quad \text{where} \quad b_{ik} \;:=\; \frac{y_i - y_k}{x_i - x_k} \quad \text{and} \quad w_{ik} \;:=\; \frac{(x_i - x_k)^2}{\sum_{i'k'} (x_{i'} - x_{k'})^2} \tag{52}$$

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

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

  The representation of β as a weighted average of case-wise slopes is

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

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

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

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

  are pairwise slopes and weights, respectively.

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

The above formulas were used and modified to produce alternative slope estimates by Gelman and Park (2008), also with the "Goal of Expressing Regressions as Comparisons that can be Understood by the General Reader" (see their Sections 1.2 and 2.2). Earlier, Wu (1986) used generalizations from pairs to tuples of size $r \geq p+1$ for the analysis of jackknife and bootstrap procedures (see his Section 3, Theorem 1). The formulas have a history in which Stigler (2001) includes Edgeworth, while Berman (1988) traces it back to a 1841 article by Jacobi written in Latin.

14 Conclusions

-
-
-
-
-

References

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

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

[3] Allison, P. D. (1995). The Impact of Random Predictors on Comparisons of Coefficients Between Models: Comment on Clogg, Petkova, and Haritou. *American Journal of Sociology* **100** (5), 1294–1305.

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

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

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

[7] 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.

[8] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26** (2), 211-252.

[9] Breiman, L. and Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review* **50** (3), 291–319.

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

[11] Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference (with discussion). *Journal of the Royal Statistical Society, Series A* **158** (3), 419-466.

[12] Clogg, C. C., Petkova, E. and Haritou, A. (1995). Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology* **100** (5), 1261–1293.

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

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

[15] Eicker, F. (1967). Limit theorems for regression with unequal and dependent errors, *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics*, Vol. 1. Berkeley: University of California Press.

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

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

[18] Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* **50**, 1029–1054.

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

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

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

[22] 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.

[23] 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.

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

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

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

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

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

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

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

A Proofs

A.1 Proofs of "Observations" in Section 3

Observation (0): Assuming constancy of the conditional distribution we obtain independence as follows:

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

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

Observation (1): This follows immediately from (5).

Observation (2): The linear case is trivial: if $\mu_0(\vec{\boldsymbol{X}})$ is linear, that is, $\mu_0(\vec{\boldsymbol{x}}) = \boldsymbol{\beta}^T\vec{\boldsymbol{x}}$ for some $\boldsymbol{\beta}$, then $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}$ irrespective of $\boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}})$ according to (5). The nonlinear case is proved as follows: For any set of points $\vec{\boldsymbol{x}}_1, ...\vec{\boldsymbol{x}}_{p+1} \in \mathbb{R}^{p+1}$ in general position and with 1 in the first coordinate, there exists a unique linear function $\boldsymbol{\beta}^T\vec{\boldsymbol{x}}$ through the values of $\mu_0(\vec{\boldsymbol{x}}_i)$. Define $\boldsymbol{P}(\mathrm{d}\vec{\boldsymbol{x}})$ by putting mass $1/(p+1)$ on each point; define the conditional distribution $\boldsymbol{P}(\mathrm{d}y\,|\,\vec{\boldsymbol{x}}_i)$ as a point mass at $y = \mu_o(\vec{\boldsymbol{x}}_i)$; this defines $\boldsymbol{P}$ such that $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}$. Now, if $\mu_0()$ is nonlinear, there exist two such sets of points with differing linear functions $\boldsymbol{\beta}_1^T\vec{\boldsymbol{x}}$ and $\boldsymbol{\beta}_2^T\vec{\boldsymbol{x}}$ to match the values of $\mu_0()$ on these two sets; by following the preceding construction we obtain $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ such that $\boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 = \boldsymbol{\beta}(\boldsymbol{P}_2)$.

A.2 Conditional Expectation of RSS

The conditional expectation of the RSS allowing for nonlinearity and heteroskedasticity:

$$\begin{aligned}
\boldsymbol{E}[\|\boldsymbol{r}\|^2|\boldsymbol{X}] &= E[\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}|\boldsymbol{X}] && (53)\\
&= \boldsymbol{E}[(\boldsymbol{X\beta}+\boldsymbol{\eta}+\boldsymbol{\epsilon})'(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{X\beta}+\boldsymbol{\eta}+\boldsymbol{\epsilon})|\boldsymbol{X}] && (54)\\
&= \boldsymbol{E}[(\boldsymbol{\eta}+\boldsymbol{\epsilon})^T(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta}+\boldsymbol{\epsilon})|\boldsymbol{X}] && (55)\\
&= \mathrm{tr}(\boldsymbol{E}[(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta}+\boldsymbol{\epsilon})(\boldsymbol{\eta}+\boldsymbol{\epsilon})^T|\boldsymbol{X}]) && (56)\\
&= \mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta\eta}^T+\boldsymbol{E}[\boldsymbol{\epsilon\epsilon}^T|\boldsymbol{X}]) && (57)\\
&= \mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta\eta}^T+\boldsymbol{D}_{\boldsymbol{\sigma}^2}) && (58)\\
&= |(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\eta}|^2+\mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{D}_{\boldsymbol{\sigma}^2}) && (59)
\end{aligned}$$

A.3 Limit of Squared Adjusted Predictors

The asymptotic limit of $\|\boldsymbol{X}_{j\hat{\bullet}}\|^2$:

$$\begin{aligned}
\frac{1}{N}\|\boldsymbol{X}_{j\hat{\bullet}}\|^2 &= \frac{1}{N}\boldsymbol{X}_j^T(\boldsymbol{I}-\boldsymbol{H}_{\text{-}j})\boldsymbol{X}_j\\
&= \frac{1}{N}\left(\boldsymbol{X}_j^T\boldsymbol{X}_j - \boldsymbol{X}_j^T\boldsymbol{H}_{\text{-}j}\boldsymbol{X}_j\right)\\
&= \frac{1}{N}X_{i,j}^2 \;-\; \left(\frac{1}{N}\sum X_{i,j}\vec{\boldsymbol{X}}_{i,\text{-}j}^T\right)\left(\sum_i \vec{\boldsymbol{X}}_{i,\text{-}j}\vec{\boldsymbol{X}}_{i,\text{-}j}^T\right)^{-1}\left(\sum_i \vec{\boldsymbol{X}}_{i,\text{-}j}X_{i,j}\right)\\
&\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[X_j^2] \;-\; \boldsymbol{E}[X_j\vec{\boldsymbol{X}}_{\text{-}j}]\boldsymbol{E}[\vec{\boldsymbol{X}}_{\text{-}j}\vec{\boldsymbol{X}}_{\text{-}j}^T]^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}_{\text{-}j}X_j]\\
&= \boldsymbol{E}[X_{j\bullet}{}^2]
\end{aligned}$$

A.4 Conventional SE

We will use the notations $\Sigma_{\vec{\boldsymbol{X}}} = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$ and $\Sigma_{\eta(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}} = \boldsymbol{E}[\eta(\vec{\boldsymbol{X}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$, and we will also need a p-dimensional normal random vector $\boldsymbol{Z}$ for the following limit in distribution:

$$\tfrac{1}{N^{1/2}} \sum_{i=1}^{N} \eta(\vec{\boldsymbol{X}}_i)\vec{\boldsymbol{X}}_i^T \;\xrightarrow{\mathcal{D}}\; \boldsymbol{Z} \;\sim\; \mathcal{N}\left(\boldsymbol{0}, \Sigma_{\eta(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}}\right) .$$

The following are the ingredients for the limiting behaviors of $\|\boldsymbol{\eta}\|^2$ and $\|\boldsymbol{H}\boldsymbol{\eta}\|^2$:

$$\begin{aligned} N\,(\boldsymbol{X}^T\boldsymbol{X})^{-1} &= \left(\frac{1}{N}\sum_{i=1}^{N} \vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i^T\right)^{-1} \\ &\xrightarrow{\boldsymbol{P}} \Sigma_{\vec{\boldsymbol{X}}}^{-1} \end{aligned}$$

$$\begin{aligned} \tfrac{1}{N}\|\boldsymbol{\eta}\|^2 &= \tfrac{1}{N}\sum_{i=1}^{N} \eta(\vec{\boldsymbol{X}}_i)^2 \\ &\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[\eta(\vec{\boldsymbol{X}})^2] \;=\; \boldsymbol{V}[\eta(\vec{\boldsymbol{X}})] \end{aligned}$$

$$\begin{aligned} \|\boldsymbol{H}\boldsymbol{\eta}\|^2 &= \boldsymbol{\eta}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta} \\ &= \left(\tfrac{1}{N^{1/2}}\sum_{i=1}^{N}\eta(\vec{\boldsymbol{X}}_i)\vec{\boldsymbol{X}}_i^T\right)\left(\tfrac{1}{N}\sum_{i=1}^{N}\vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i^T\right)^{-1}\left(\tfrac{1}{N^{1/2}}\sum_{i=1}^{N}\eta(\vec{\boldsymbol{X}}_i)\vec{\boldsymbol{X}}_i\right) \\ &\xrightarrow{\mathcal{D}} \boldsymbol{Z}^T\Sigma_{\vec{\boldsymbol{X}}}^{-1}\boldsymbol{Z} \end{aligned}$$

$$\tfrac{1}{N}\|\boldsymbol{H}\boldsymbol{\eta}\|^2 \xrightarrow{\boldsymbol{P}} 0$$

For large N and fixed p, as $N/(N-p-1) \to 1$ we have

$$N\,\frac{\frac{1}{N-p-1}\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\eta}\|^2}{\|X_{j\bullet}\|^2} \;\xrightarrow{\boldsymbol{P}}\; \frac{\boldsymbol{E}[\eta(\vec{\boldsymbol{X}})^2]}{\boldsymbol{E}[X_{j\bullet}{}^2]} . \qquad (60)$$

A.5 Asymptotic Normality in Terms of Adjustment

We gave the asymptotic limit of the conditional bias in vectorized form after (20)-(22). Here we derive the equivalent element-wise limit using adjustment to show (**??**)-(**??**). The variance of the conditional bias is the marginal inflator of SE.

$$N^{1/2}(\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}] - \beta_j) \;=\; N^{1/2}\frac{\langle X_{j\bullet}, \boldsymbol{\eta}\rangle}{\|X_{j\bullet}\|^2} \;=\; \frac{\frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{\eta} - \frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{H}_{-j}\boldsymbol{\eta}}{\frac{1}{N}\|X_{j\bullet}\|^2}$$

$$\begin{aligned}
\frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{H}_{-j}\boldsymbol{\eta} &= \frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^T\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^T\boldsymbol{\eta} \\
&= \left(\frac{1}{N}\sum_i X_{i,j}\vec{\boldsymbol{X}}_{i,-j}^T\right)\left(\frac{1}{N}\sum_i \vec{\boldsymbol{X}}_{i,-j}\vec{\boldsymbol{X}}_{i,-j}^T\right)^{-1}\left(\frac{1}{N^{1/2}}\sum_i \vec{\boldsymbol{X}}_{i,-j}\eta(\vec{\boldsymbol{X}}_i)\right) \\
&\overset{\mathcal{D}}{\approx} \boldsymbol{E}[X_j\vec{\boldsymbol{X}}_{-j}]\boldsymbol{E}[\vec{\boldsymbol{X}}_{-j}\vec{\boldsymbol{X}}_{-j}^T]\left(\frac{1}{N^{1/2}}\sum_i \vec{\boldsymbol{X}}_{i,-j}\eta(\vec{\boldsymbol{X}}_i)\right) \\
&= \boldsymbol{\beta}_{j\cdot}^T\left(\frac{1}{N^{1/2}}\sum_i \vec{\boldsymbol{X}}_{i,-j}\eta(\vec{\boldsymbol{X}}_i)\right) \\
&= \frac{1}{N^{1/2}}\sum_i(\boldsymbol{\beta}_{j\cdot}^T\vec{\boldsymbol{X}}_{i,-j})\eta(\vec{\boldsymbol{X}}_i)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{N^{1/2}}\left(\boldsymbol{X}_j^T\boldsymbol{\eta} - \boldsymbol{X}_j^T\boldsymbol{H}_{-j}\boldsymbol{\eta}\right) &\overset{\mathcal{D}}{\approx} \frac{1}{N^{1/2}}\sum_i\left(X_{i,j} - \boldsymbol{\beta}_{j\cdot}^T\vec{\boldsymbol{X}}_{i,-j}\right)\eta(\vec{\boldsymbol{X}}_i) \\
&\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \boldsymbol{V}[(X_j - \boldsymbol{\beta}_{j\cdot}^T\vec{\boldsymbol{X}}_{-j})\eta(\vec{\boldsymbol{X}})]\right) \\
&= \mathcal{N}\left(0, \boldsymbol{V}[X_{j\bullet}\eta(\vec{\boldsymbol{X}})]\right)
\end{aligned}$$

$$N^{1/2}(\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}] - \beta_j) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\boldsymbol{V}[X_{j\bullet}\eta(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}\right)$$

A.6 A Family of Nonlinearities with Extreme *RAV*

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

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

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

$$\eta_t(X) \;=\; \frac{1_{[|X|>t]} - p(t)}{\sqrt{p(t)(1-p(t))}}\,, \qquad \textit{where} \quad p(t) = \boldsymbol{P}[|X| > t]\,, \tag{61}$$

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

$\lim_{t\uparrow\infty} \boldsymbol{RAV}(\eta_t^2) = \infty$;

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

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

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

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

$$\sigma_t^2(X) \;=\; \frac{(1_{[|X|>t]} - p(t))^2}{p(t)(1-p(t))}\,, \qquad \textit{where} \qquad p(t) = \boldsymbol{P}[|X| > t]\,, \tag{62}$$

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

$\lim_{t\uparrow\infty} \boldsymbol{RAV}(\sigma_t^2) = \infty$;

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

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

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

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

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

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

A.7 Details for the heteroskedasticity and nonlinearity examples

We write X instead of $X_{j\bullet}$ and assume it has a standard normal distribution, $X \sim N(0,1)$, whose density will be denoted by $\phi(x)$. In Figure 5 the base function is, up to scale, as follows:

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

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

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

Accordingly we obtain the following moments:

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

and hence

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

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

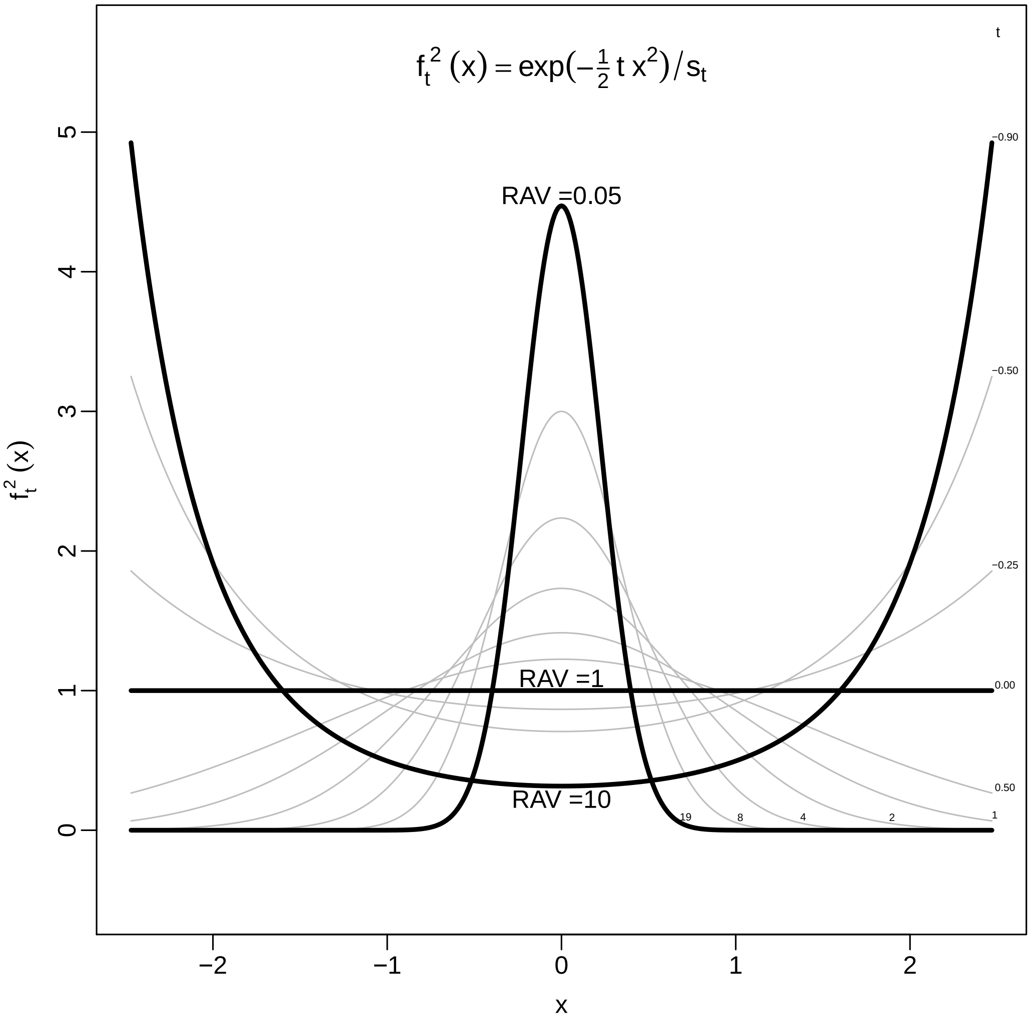

Figure 5: *A family of functions $f_t^2(x)$ that can be interpreted as heteroskedasticities $\sigma^2(X_{j\bullet})$, squared nonlinearities $\eta^2(X_{j\bullet})$, or conditional MSEs $m^2(X_{j\bullet})$: The family interpolates* ***RAV*** *from* 0 *to* ∞ *for* $X_{j\bullet} \sim N(0,1)$.
RAV $= \infty$ *is approached as $f_t^2(x)$ bends ever more strongly in the tails of the x-distribution.*
RAV $= 0$ *is approached by an ever stronger spike in the center of the x-distribution.*
(See Appendix A.7 for details.)

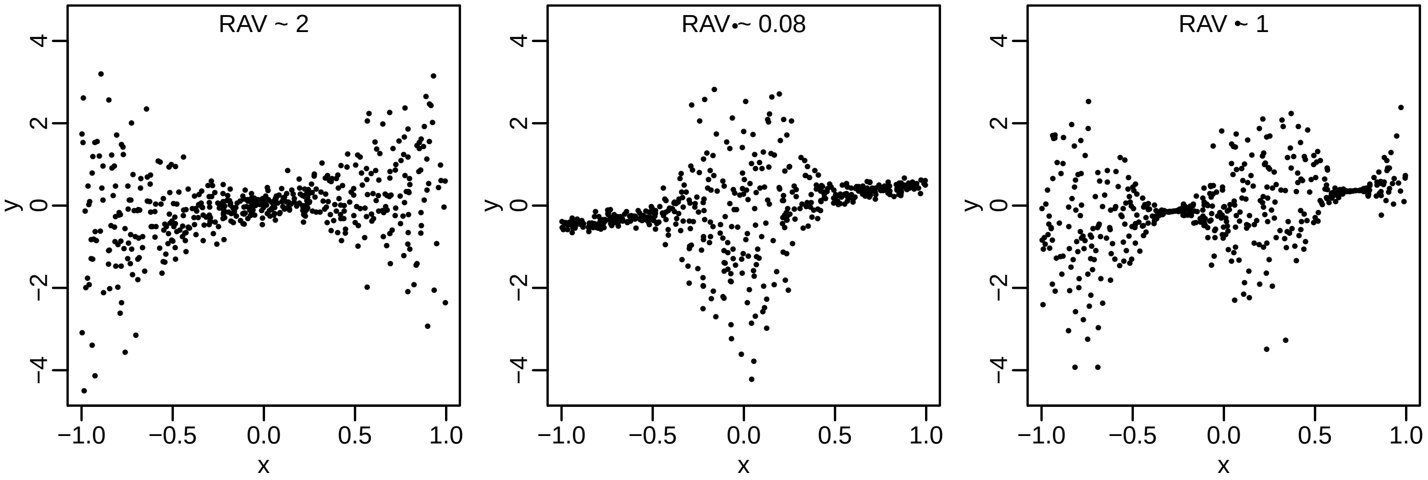

Figure 6: *The effect of heteroskedasticity on the sampling variability of slope estimates: The question is how the misinterpretation of the heteroskedasticities as homoskedastic affects statistical inference.*
Left: High error variance in the tails of the predictor distribution elevates the true sampling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

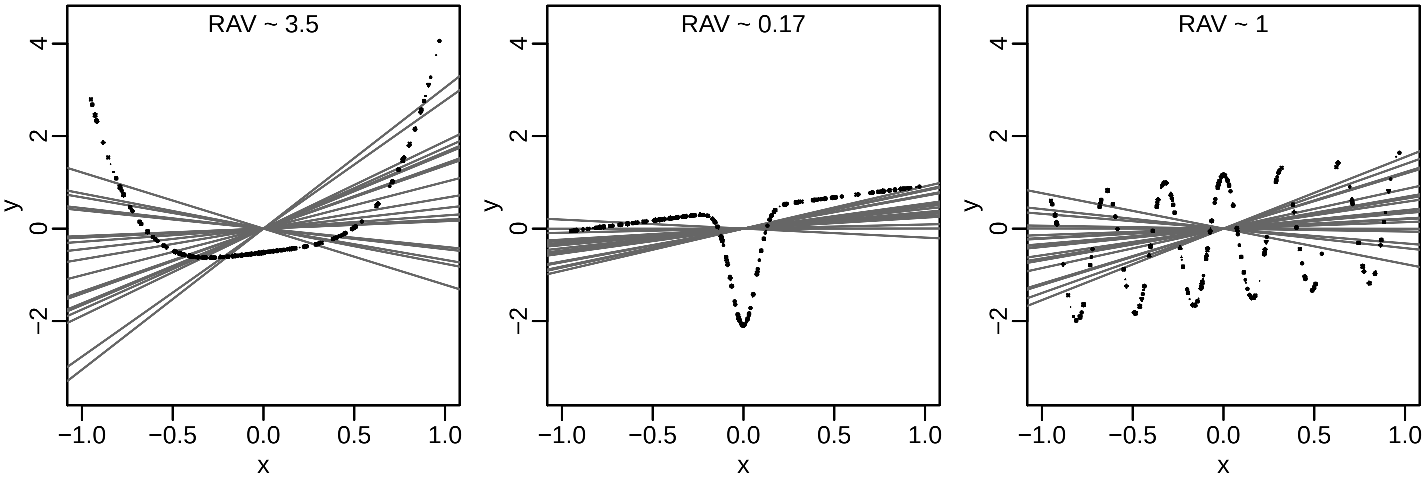

Figure 7: *The effect of nonlinearities on the sampling variability of slope estimates: The three plots show three different error-free nonlinearities; each plot shows for one nonlinearity 20 overplotted datasets of size $N = 10$ and their fitted lines through the origin. The question is how the misinterpretation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

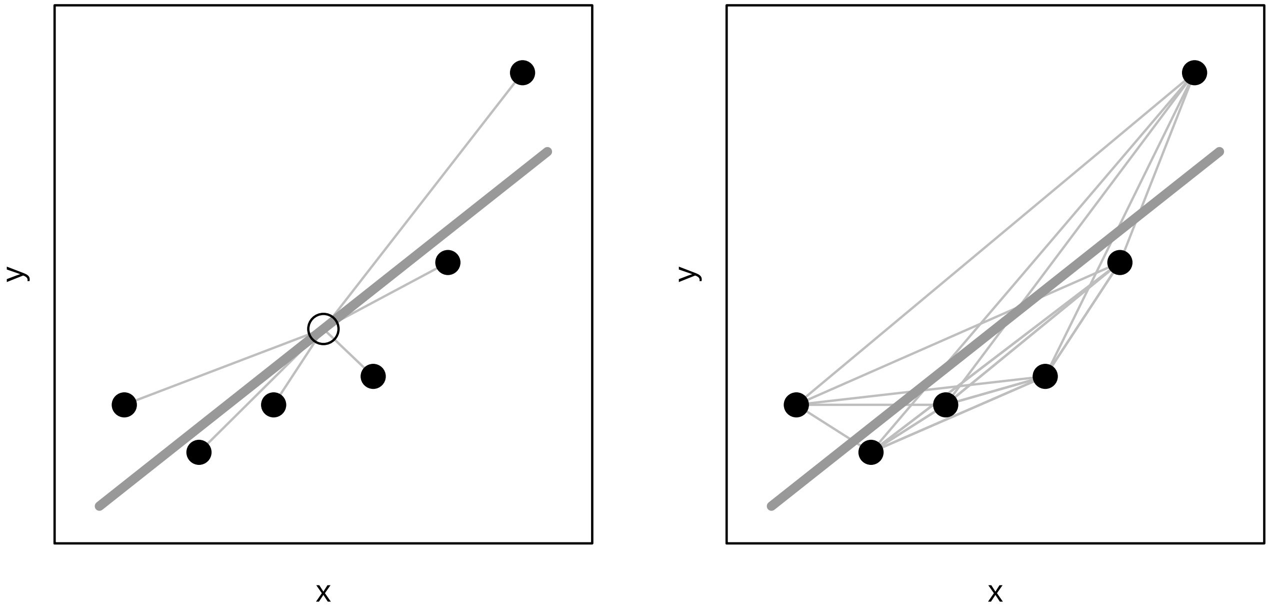

Figure 8: *Case-wise and pairwise average weighted slopes illustrated: Both plots show the same six points (the "cases") as well as the LS line fitted to them (fat gray). The left hand plot shows the case-wise slopes from the mean point (open circle) to the six cases, while the right hand plot shows the pairwise slopes between all 15 pairs of cases. The LS slope is a weighted average of the case-wise slopes on the left according to* (51)*, and of the pairwise slopes on the right according to* (52)*.*