A Conspiracy of Random X and Model Violation against Classical Inference in Linear Regression

A. Buja, R. Berk, L. Brown, E. George, M. Traskin, K. Zhang, L. Zhao
The Wharton School, University of Pennsylvania

April 19, 2013

Abstract

Following the econometric literature on model misspecification, we examine statistical inference for linear regression coefficients β_j when the predictors are random and the linear model assumptions of first and/or second order are violated: $\boldsymbol{E}[Y|X_1, ..., X_p]$ is not linear in the predictors and/or $\boldsymbol{V}[Y|X_1, ..., X_p]$ is not constant. Such inference is meaningful if the linear model is seen as a useful approximation rather than part of a generative truth.

A difficulty well-known in econometrics is that the required standard errors under random predictors and model violations can be very different from the conventional standard errors that are valid when the linear model is correct. The difference stems from a synergistic effect between model violations and randomness of the predictors. We show that asymptotically the ratios between correct and conventional standard errors can range between infinity and zero. Also, these ratios vary between predictors within the same multiple regression.

This difficulty has consequences for statistics: It entails the breakdown of the classical ancillarity argument for predictors. If the assumptions of a generative regression model are violated, then the ancillarity argument for the predictors no longer holds, treating predictors as fixed is no longer valid, and standard inferences may lose their significance and confidence guarantees.

The standard econometric solution for consistent inference under misspecification and random predictors is based on the "sandwich estimator" of the covariance matrix of $\hat{\boldsymbol{\beta}}$. A plausible alternative is the paired bootstrap which resamples predictors and response jointly. Discrepancies between conventional and bootstrap standard errors can be used as diagnostics for predictor-specific model violations, in analogy to econometric misspecification tests. The good news is that when model violations are sufficiently strong to invalidate conventional linear inference, their nature tends to be visible in graphical diagnostics.

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

1 Introduction

Classical inference for linear regression is conditional on the observed predictors. This means that in a linear model

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

the standard errors $\mathrm{SE}(\hat{\beta}_j)$ only account for the variability in the error vector $\boldsymbol{\epsilon}$ but not in the predictor data $\boldsymbol{X}$. In effect, the predictor data are treated as fixed known constants even when they are samples, as they are in observational data. The justification for conditioning on random $\boldsymbol{X}$ is that the predictor distribution is ancillary

for the parameters $\boldsymbol{\beta}$ and σ of the model, hence conditioning on $\boldsymbol{X}$ produces valid frequentist inference for these parameters (Cox and Hinkley 1974, Example 2.27).

A problem with the ancillarity argument is that it holds only when the linear model is correct. In practice, of course, whether a model is correct is never known, yet it is fitted just the same and also richly interpreted with statistical inference that assumes the truth of the model. An awareness of potential problems with such inference is occasionally handled with a reference to "robustness of validity", meaning that conventional t- and F-tests are conservative (though not efficient) when the error distribution is heavier-tailed than normal. This argument, however, does not address model failure from non-linearity or heteroskedasticity. When first and/or second order assumptions are violated, there may arise pernicious problems for statistical inference because these violations can interact with the randomness of the predictors and invalidate conventional standard errors derived from $\boldsymbol{X}$-conditional linear models theory. As we will see, this interaction usually inflates and occasionally deflates the sampling variability of slope estimates relative to conventional $\boldsymbol{X}$-conditional standard errors.

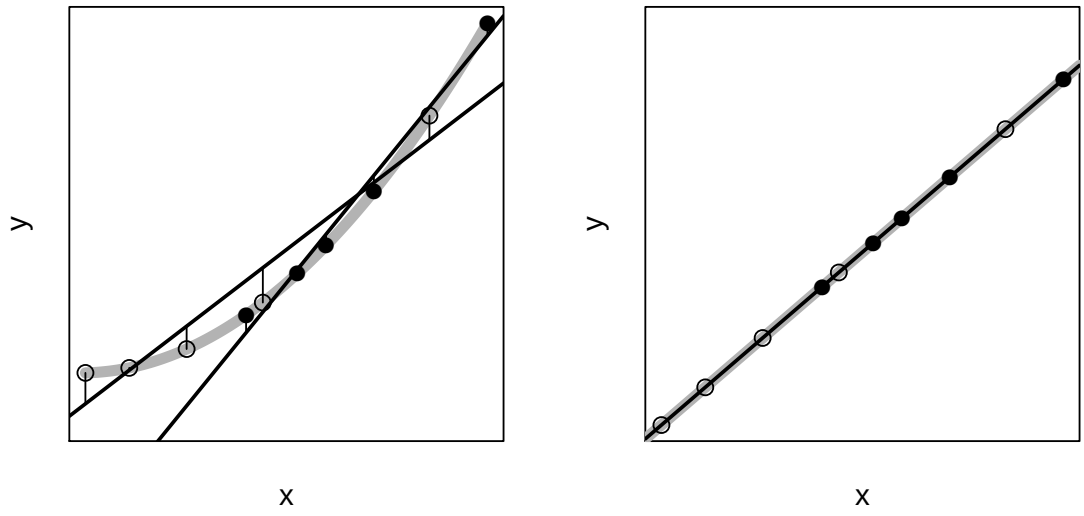

Figure 1: *Single Predictor, error-less Data: The filled and the open circles represent two "datasets" from the same population. The x-values are random; the y-values are a deterministic function of x: $y = \mu(x)$ (shown in gray).*
Left: The true response $\mu(x)$ (gray curve) is nonlinear; the open and the filled circles have different LS lines (black lines).
Right: The true response $\mu(x)$ (gray line) is linear; the open and the filled circless have the same LS line (black overlayed on gray).

The nature of the interaction between model violations and randomness of the predictors is most drastically illustrated with the artificial example of an error-free nonlinearity, that is, $y = \mu(x)$ is a deterministic but nonlinear function of a (single) random predictor x, as in the left hand plot of Figure 1. It shows two error-free (x, y) "datasets" of size $N = 5$ each, marked by open circles ("white data") and filled circles ("black data"). The datasets are clearly shifted against each other, the white data being more to the left, the black data more to the right, even though both are i.i.d. random samples from the same predictor distribution. The relative shift implies that when a straight line is fitted to the "response" in both datasets the white dataset will "see" a smaller slope than the black dataset. This difference is solely due to the random differences between the two predictor samples. — This effect vanishes when the response function $\mu(x)$ is linear, as depicted in the right hand plot of Figure 1 for the same two sets of predictor values: obviously both sets "see" the same slope. Thus it is the combination

of randomness of the x-values and nonlinearity of the response function $\mu(x)$ that conspires to create sampling variability in LS estimates of slopes (and intercepts as well). This additional source of sampling variability is separate from the variability analyzed in $\boldsymbol{X}$-conditional linear models theory, which is error in the response. Furthermore, the additional variability is of the same order $1/\sqrt{N}$ as the conventional error-based variability.

In the econometric literature on misspecification it is known that standard errors can be invalidated not only by nonlinearity but by heteroskedasticity as well. Heteroscedasticity's effect on sampling variability is somewhat less transparent than nonlinearity's, yet in econometrics it is heteroscedasticity that appears to be more frequently debated. In the present context the relative importance of the two effects is not the issue; rather, it is the sheer fact that statistical inference is affected by an interaction of first and second order model violations with predictor randomness.

An immediate consequence for classical statistics is that for observational data $\boldsymbol{X}$-conditional inference must be considered as suspect, and so must the ancillarity argument that justifies it. The ancillarity argument for regression has been unquestioned by statisticians for many decades, and this acceptance may be partly supported by a fallacious intuitition. The fallacy may run along the following lines: One starts from the premise that if the data do not satisfy the first and second order assumptions for the model at hand, they will do so very likely for some larger linear model. In that "correct" model the ancillarity argument applies, hence it applies in the smaller model. This very last step is of course the fallacy: ancillarity for the predictors can only be argued in the correct model, not its incorrect submodels.

A further consequence for classical statistics is that if $\boldsymbol{X}$-conditional inference is to be used, then checking first and second order correctness of the chosen model is essential. To address this need we will propose some simple diagnostics that provide indications of model failure in a predictor-specific manner. Such diagnostics have of course precursors in econometric misspecification tests. Our diagnostics will add a new level of interpretability to these established techniques [??? make more specific].

Standard errors that capture the variability generated by the interaction of random predictors and model violations have existed for quite some time: asymptotically correct standard errors can be obtained from so-called "sandwich estimators" of the covariance matrix of the estimates (White 1980a, 1980b, 1981, 1982). These have generated a considerable literature and have been applied to every conceivable type of parametric regression. They are better known for being "heteroskedasticity consistent" (White 1980b) than nonlinearity consistent (White 1980a).

Another and quite obvious approach to standard errors under random predictors and model violations is bootstrapping $(\vec{\boldsymbol{X}}_i, Y_i)$ pairs. It captures the joint, unconditional sampling variability of the response *and* the predictors without making model assumptions other than i.i.d. sampling. There exists a literature that shows this bootstrap to be asymptotically correct for linear regression under very general conditions, even if the linear model is incorrect; see for example Freedman's (1981) analysis of the "correlation model", and Mammen's (1993) analysis for increasing p. The paired bootstrap, however, has been a source of unease to some statisticians because it violates conditionality on the predictors. As we have seen above, such unease is ill-founded because conditionality relies on the ancillarity argument for predictors, which in turn assumes the correctness of the model. Such assumption creates exposure to potentially flawed statistical inference.

For linear models applied to observational data it should be recommended that all inference be accompanied by standard errors obtained from either sandwich estimators or a paired bootstrap simulation, if only as a diagnostic. If data analysts observe sizable discrepancies between conventional and bootstrap standard errors, it should alert them to potential problems on two levels: (1) the existence of first and/or second order model incorrectness and consequently a need to diagnose the nature of nonlinearities and/or heteroskedasticities, and (2) a need to use the

bootstrap standard errors for statistical inference if these are wider.

The reader may object at this point that something is amiss with this argument: If the functional form of the model is incorrect and if there is indeed a nonlinear association between predictors and response, what is the meaning of a slope and its standard error? Shouldn't the nonlinearities be taken care of first, for example, by fitting an additive model (Buja, Hastie, and Tibshirani 1989; Hastie and Tibshirani 1990) and/or transforming the response (Box and Cox 1964)? These arguments are of course valid: competent data analysts would discover critical nonlinearities with residual analysis, non-parametric fitting, nonlinear transformation of the response, and other methods of model building. Yet, the counterargument is that even when a linear model for observational data is first order incorrect, it can still be *theoretically meaningful* and *practically useful* (viz Box' (1979, second section heading) famous dictum). Both assertions can be made precise:

- Theoretically, a fitted linear model can always be interpreted as estimating the best linear approximation $\beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ to the conditional expectation $\boldsymbol{E}[Y|X_1, ..., X_p]$ at the population. This view is in line with machine learning approaches which investigate best performance within a functional form without assuming it to contain the truth. The difference of our focus to that in machine learning is that we are not concerned with prediction but with valid inference for the parameters of the fitted equation.
- Practically, a fitted linear model, even if not strictly correct, can give information about the direction of the association — positive or negative — between Y and X_j in the presence of the other predictors. Also, it is possible to give meaning to LS slopes as weighted averages of "observed slopes", and this carries over to populations without linear model assumptions. Because of its intuitive appeal we discuss this interpretation in Appendix A.

In conclusion, the problem of inference for "incorrect" linear models is not misguided; it can be both meaningful and useful.

The idea that models are approximations, not generative truths, is found in a literature too large to list. Aspects of approximation are explicit in work on robustness, on misspecified models, and on nonparametric fitting. Most famously the idea of models as approximations is implied by Box (1979), albeit in more catchy language. On this matter we follow Cox' (1995) opinion that "it does not seem helpful to say that all models are wrong. The very word model implies simplification and idealization," which implies approximation. The theme runs through the discussion generated by Chatfield (1995), which is where we found Cox' comment. — Among the several meanings of "approximation" as they relate to statistical models, the only one we have in mind is that of approximating the conditional expectation $\boldsymbol{E}[Y|X_1, ..., X_p]$ of a numerically interpreted response by a linear function $\beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ of the predictors. The possible causes of nonlinearity include incorrect functional form, omission of predictors, and measurement errors in predictors. Whatever the causes, known or unknown, we assume nevertheless that interest is in $\boldsymbol{E}[Y|X_1, ..., X_p]$ through inference for the coefficients of the best linear approximant $\beta_0 + \beta_1 X_1 + ... + \beta_p X_p$.

This article continuous as follows: Section 2 shows a data example in which the conventional and bootstrap standard errors of some slopes are off by as much as a factor 2. Section 3 introduces notation for populations and samples, LS approximations, adjustment operations, nonlinearities, and decompositions of responses. Section 4 describes the sampling variation caused by random predictors and nonlinearities. Section 5 compares 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 6 shows that asymptotically there is no limit to which inflation of the unconditional standard error can exceed inflation of the conventional standard error. Appendix A gives intuitive meaning to LS slopes in terms of weighted average slopes found in the data.

2 The Problem Illustrated with

We will use the well-known Boston Housing data for illustration, but we must begin with a caveat: Theseq data (Harrison and Rubinfeld 1978) are an enumeration of the census tracts in the larger Boston area. They do **not** form a random sample; at a minimum, stochastic modeling would have to include consideration of spatial correlation. Yet, these data have been used to illustrated regression methods at least since the regression diagnostics book by Belseley, Kuh and Welsch (1980). In what follows we will treat the data as if they were a random sample to illustrate the issue of conditional versus unconditional statistical inference in linear regression. Because of pecularities in the associations among the 14 variables in these data, they are uniquely suited to illustrate the phenomenon we describe here.

The following table shows in the first three numeric columns the results of a linear regression of `MEDV` (median value of owner occupied homes, standardized) on thirteen predictors (all standardized). As we can see, the predictors `RM` (average number of rooms per dwelling) and `LSTAT` (percent of lower status population) are the strongest in terms of t-ratios.

	Slope.est	SE.conv	t.val		SE.boot	SE.boot/SE.conv	
CRIM	-0.099	0.031	-3.261		0.033	1.074	
ZN	0.121	0.035	3.508		0.035	1.004	
INDUS	0.017	0.046	0.382		0.038	0.843	
CHAS	0.074	0.024	3.152		0.036	1.503	<?
NOX	-0.224	0.048	-4.687		0.048	1.003	
RM	0.290	0.032	9.149	<<	0.065	2.049	<<
AGE	0.002	0.040	0.044		0.050	1.236	
DIS	-0.344	0.045	-7.598	<	0.048	1.068	
RAD	0.288	0.062	4.620		0.060	0.958	
TAX	-0.233	0.068	-3.409		0.051	0.740	
PTRATIO	-0.218	0.031	-7.126	<	0.026	0.865	
B	0.092	0.026	3.467		0.027	1.036	
LSTAT	-0.413	0.039	-10.558	<<	0.078	1.995	<<

Important for us is the comparison of the conventional standard errors (`SE.conv`, conditional on $\boldsymbol{X}$) and the paired bootstrap standard errors (`SE.boot`, sampling (x, y) pairs, not residuals) in the second and the fourth numeric column, respectively. The last column shows their ratio. We see that for the two strongest predictors, `LSTAT` and `RM`, the bootstrap standard errors are about twice the size of the conventional standard errors. As a consequence, these predictors seem to have their t-statistics of -10.558 and 9.149 inflated by about a factor of two. If so, they would have to cede their first and second ranks to the predictors `DIS` and `PTRATIO` whose bootstrap standard errors are in rough agreement with the traditional standard errors and who exhibit t-statistics of -7.598 and -7.126. So we must ask how a discrepancy can arise between bootstrap and traditional standard errors.

A pointer toward an answer is given in Figure 2: We are first shown residuals plotted against all predictors, augmented with smooths. The smooths indicate that the predictors `LSTAT` and `RM` are those that most strongly suggest nonlinear effects. Even more convincingly, a second series of plots shows the results of an additive model fit, $Y \sim \phi_1(X_1) + ... + \phi_p(X_p)$, augmented with bootstrap bands. The transformations of the predictors `LSTAT` and `RM` suggest most strongly a nonlinear effect once again. Other predictors may also have nonlinear associations, but `LSTAT` and `RM` are the two most egregious cases.

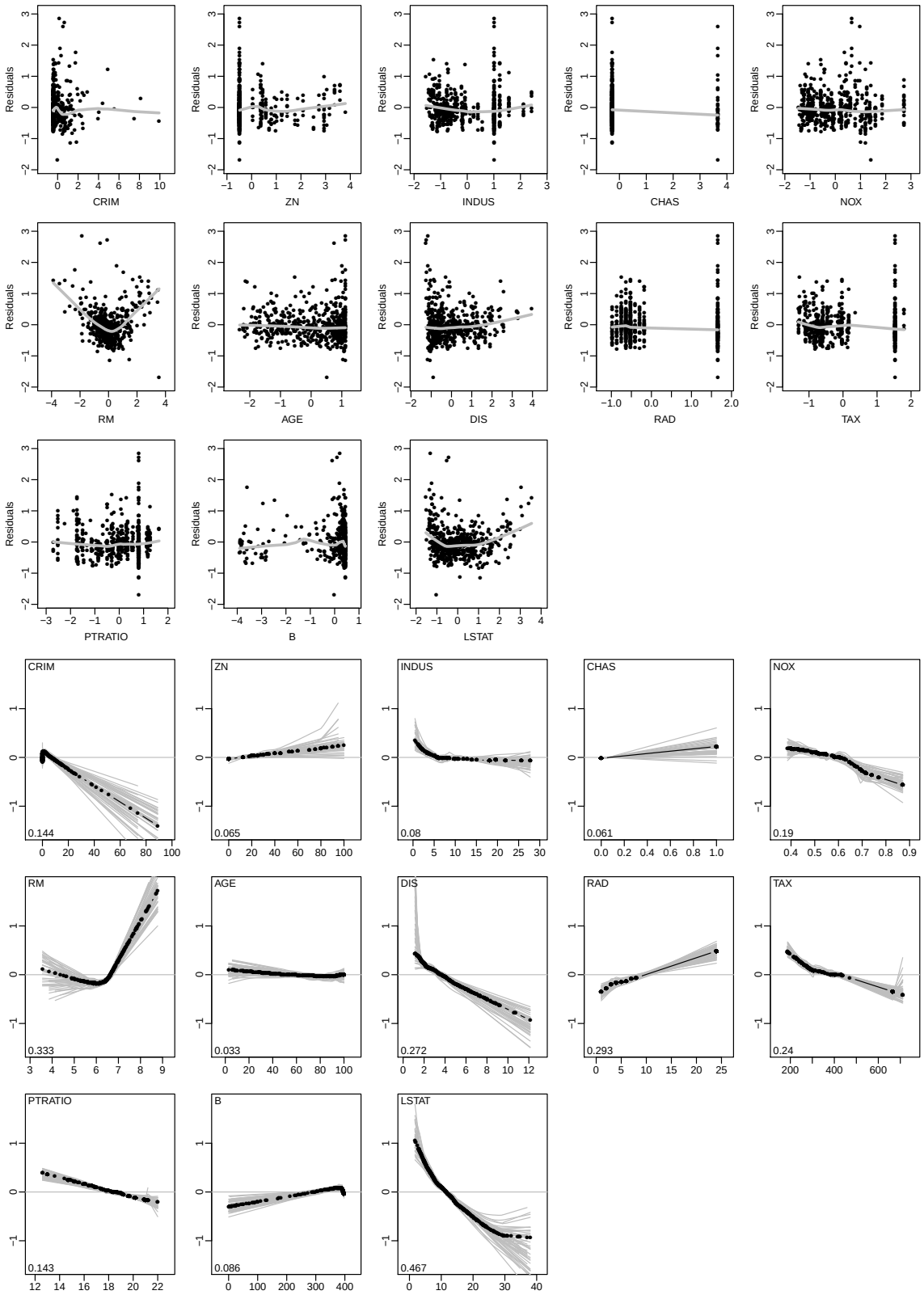

Figure 2: *Boston Housing data: The strongest two predictors,* `RM` *and* `LSTAT`*, show substantial nonlinearities.*
Top half: Residuals versus all predictors augmented with lowess smooths.
Bottom half: ACE transformations for an additive model; the response is standardized but otherwise untransformed. The bottom left corners show the standard deviations of the transforms (= analogs of $|\hat{\beta}_j|$ *when the* $\mathrm{sd}(\boldsymbol{x}_j) = 1$*).*

3 Observational Regression Data: Estimation and its Target

Following Freedman (1981) and Mammen (1993), we introduce a framework in which linear LS is applied to data that are not described by a generative linear model but only by an i.i.d. sampling assumption. The framework therefore consists of (1) a largely arbitrary multivariate population in which one variable is singled out as the response, (2) the associated decomposition of the population response into a linear LS approximation, a nonlinearity, and error, (3) samples from the population as well as LS estimates obtained from the samples.

3.1 Quantities associated with the population

- The population will be described by random variables X_1, ..., X_p and Y. At this point the only assumption is that these variables have a nontrivial joint distribution $\boldsymbol{P} = \boldsymbol{P}(dx_1, ..., dx_p, dy)$ in the sense that second moments exist and the covariance matrix is of full rank.

- 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. We write any function $f(X_1, ..., X_p)$ equivalently as $f(\vec{\boldsymbol{X}})$ as the prepended constant 1 is irrelevant.

- The "true response surface" $\mu(\vec{\boldsymbol{X}})$ is the conditional mean given the predictors $\vec{\boldsymbol{X}}$:

$$\mu(\vec{\boldsymbol{X}}) \;=\; \boldsymbol{E}[\, Y | \vec{\boldsymbol{X}}\,]\,. \tag{2}$$

  No assumption of linearity as a function of $\vec{\boldsymbol{X}}$ is made.

- The **population linear LS approximation** to Y is the affine function defined by

$$\boldsymbol{\beta}^T \vec{\boldsymbol{X}} \;=\; \beta_0 + \beta_1 X_1 + ... + \beta_p X_p\,, \tag{3}$$

  where the population intercept and slopes are collected in the $(p+1)$-vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ and $\boldsymbol{\beta}$ is defined by the following equivalent expressions:

$$\begin{aligned}
\boldsymbol{\beta}(\boldsymbol{P}) \;&=\; \text{argmin}_{\boldsymbol{\beta}}\, \boldsymbol{E}[\,(Y - \boldsymbol{\beta}^T \vec{\boldsymbol{X}})^2\,] \;=\; \text{argmin}_{\boldsymbol{\beta}}\, \left(-2\boldsymbol{\beta}^T \boldsymbol{E}[Y\vec{\boldsymbol{X}}] + \boldsymbol{\beta}^T \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\boldsymbol{\beta}\right) && (4)\\
&=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}\, \boldsymbol{E}[\,Y\vec{\boldsymbol{X}}\,] && (5)\\
&=\; \text{argmin}_{\boldsymbol{\beta}}\, \boldsymbol{E}[\,(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}^T \vec{\boldsymbol{X}})^2\,] && (6)\\
&=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}\, \boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\,]\,. && (7)
\end{aligned}$$

  Equalities (5)-(7) follow from the normal equations for the population linear LS problem in (4):

$$\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]\,\boldsymbol{\beta} \;=\; \boldsymbol{E}[\,Y\vec{\boldsymbol{X}}\,] \;=\; \boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\,]\,. \tag{8}$$

  The vector $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ is a population parameter and the target of estimation. Unlike in a linear model, it only describes the best linear approximation to $\mu(\vec{\boldsymbol{X}})$, not $\mu(\vec{\boldsymbol{X}})$ itself. There is no generative model assumed here, hence $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ is just a statistical functional defined on multivariate distributions $\boldsymbol{P} = \boldsymbol{P}(dx_1, ..., dx_p, dy)$.

- **Decompositions** of the response and response function:

Response:	Y			
Response function:	$\mu(\vec{\boldsymbol{X}})$	$= \boldsymbol{E}[\,Y\,\vert\,\vec{\boldsymbol{X}}]$		
Population LS approximation:	$\boldsymbol{\beta}^T\vec{\boldsymbol{X}}$	(see (4)-(7))		
Nonlinearity, model bias:	$\eta(\vec{\boldsymbol{X}})$	$= \mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}^T\vec{\boldsymbol{X}}$		(9)
Error, noise:	ϵ	$= Y - \mu(\vec{\boldsymbol{X}})$		(10)
Population residual:	ξ	$= Y - \boldsymbol{\beta}^T\vec{\boldsymbol{X}}$	$= \eta(\vec{\boldsymbol{X}}) + \epsilon$	(11)
Error variance:	$\sigma^2(\vec{\boldsymbol{X}})$	$= \boldsymbol{V}[Y\vert\vec{\boldsymbol{X}}]$	$= \boldsymbol{E}[\,\epsilon^2\vert\vec{\boldsymbol{X}}]$	(12)
Conditional MSE:	$\rho^2(\vec{\boldsymbol{X}})$	$= \boldsymbol{E}[\,\xi^2\vert\vec{\boldsymbol{X}}]$	$= \eta^2(\vec{\boldsymbol{X}}) + \sigma^2(\vec{\boldsymbol{X}})$	(13)

  The first three items are recapitulations, and the rest are definitions. Crucial are the *nonlinearity* or *model bias* function $\eta(\vec{\boldsymbol{X}})$ to permit first order incorrectness, and the *error variance* function $\sigma^2(\vec{\boldsymbol{X}})$ to permit second order model incorrectness. For lack of a better name we call ξ the *population residual*. The error ϵ satisfies $\boldsymbol{E}[\,\epsilon\,\vert\vec{\boldsymbol{X}}] = 0$.
  [Caution: Freedman (1981) and Mammen (1993) use "ϵ" to mean $\xi = \eta + \epsilon$ in our notation. They do not divide the population residual into error and nonlinearity.]

- **Orthogonalities/decorrelation** at the population: All of ξ, $\eta(\vec{\boldsymbol{X}})$ and ϵ are orthogonal to the predictors:

$$\boldsymbol{E}[\,\xi\vec{\boldsymbol{X}}] \;=\; \boldsymbol{E}[\,\eta(\vec{\boldsymbol{X}})\,\vec{\boldsymbol{X}}] \;=\; \boldsymbol{E}[\,\epsilon\,\vec{\boldsymbol{X}}] \;=\; \boldsymbol{0}\,. \tag{14}$$

  The reason for ξ and $\eta(\vec{\boldsymbol{X}})$ is that they are residuals in the population linear LS regressions of Y and $\mu(\vec{\boldsymbol{X}})$, respectively, on $\vec{\boldsymbol{X}}$. The error ϵ is orthogonal to $\vec{\boldsymbol{X}}$ because $\boldsymbol{E}[\,\epsilon\,\vert\vec{\boldsymbol{X}}] = 0$. — As a consequence of the inclusion of an intercept entry "1" in $\vec{\boldsymbol{X}}$, orthogonality of ξ and $\eta(\vec{\boldsymbol{X}})$ to $\vec{\boldsymbol{X}}$ entails centering, whereas for ϵ it follows from $\boldsymbol{E}[\,\epsilon\,\vert\vec{\boldsymbol{X}}] = 0$:

$$\boldsymbol{E}[\,\xi\,] \;=\; \boldsymbol{E}[\,\eta(\vec{\boldsymbol{X}})\,] \;=\; \boldsymbol{E}[\,\epsilon\,] \;=\; 0\,. \tag{15}$$

  The error ϵ satisfies much stronger orthogonalities: For any functions $f(\vec{\boldsymbol{X}})$ and $g(\vec{\boldsymbol{X}})$ (with suitable moments), $\epsilon f(\vec{\boldsymbol{X}})$ is orthogonal to $g(\vec{\boldsymbol{X}})$. This means $\boldsymbol{E}[(\epsilon f(\vec{\boldsymbol{X}}))g(\vec{\boldsymbol{X}})] = 0$, which is an immediate consequence of $\boldsymbol{E}[\epsilon|\vec{\boldsymbol{X}}] = 0$. Yet ϵ is not generally independent of $\vec{\boldsymbol{X}}$.

- **Adjustment formulas for the population**: We introduce the population version of "adjusted" predictors in order to express the multiple regression coefficient β_j as a simple regression coefficient. The adjusted predictor $X_{j\bullet}$ is defined as the residual of the regression of X_j (used as the response) on all other predictors. To this end we collect all other predictors by writing $\vec{\boldsymbol{X}}_{\setminus j} = (1, X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)^T$ and defining the p-vector $\boldsymbol{\beta}_{j\bullet}$ by

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

  The adjusted predictor is

$$X_{j\bullet} \;=\; X_j - \boldsymbol{\beta}_{j\bullet}^T\vec{\boldsymbol{X}}_{\setminus j}\,. \tag{16}$$

  The representation of β_j as a simple regression coefficient is

$$\beta_j \;=\; \frac{\boldsymbol{E}[YX_{j\bullet}]}{\boldsymbol{E}[{X_{j\bullet}}^2]} \;=\; \frac{\boldsymbol{E}[\mu(\vec{\boldsymbol{X}})X_{j\bullet}]}{\boldsymbol{E}[{X_{j\bullet}}^2]}\,. \tag{17}$$

  This is a componentwise alternative to (5) above.

3.2 Notation for observational datasets

- When we use the term "observational data" we actually mean "cross-sectional data" consisting of i.i.d. cases $(X_{i,1}, ..., X_{i,p}, Y_i)$ drawn from the multivariate distribution $\boldsymbol{P}(dx_1, ..., dx_p, dy)$ $(i = 1, 2, ..., N)$. 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:

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

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

$$\boldsymbol{\mu} = \begin{bmatrix} \mu(\vec{\boldsymbol{X}}_1) \\ \cdot\cdot \\ \cdot\cdot \\ \mu(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta(\vec{\boldsymbol{X}}_1) \\ \cdot\cdot \\ \cdot\cdot \\ \eta(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \cdot\cdot \\ \cdot\cdot \\ \epsilon_N \end{bmatrix}, \quad \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \cdot\cdot \\ \cdot\cdot \\ \xi_N \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma(\vec{\boldsymbol{X}}_1) \\ \cdot\cdot \\ \cdot\cdot \\ \sigma(\vec{\boldsymbol{X}}_N) \end{bmatrix}. \tag{18}$$

- The definitions (11) of ξ, (10) of ϵ, and (9) of $\eta(\vec{\boldsymbol{X}})$ translate to vectorized forms:

$$\boldsymbol{\xi} \;=\; \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} \;=\; \boldsymbol{Y} - \boldsymbol{\mu}, \qquad \boldsymbol{\eta} \;=\; \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}. \tag{19}$$

- It is important to keep in mind the distinction between population and sample properties, in particular:
  - The $(p+1)$-vector $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ is *not* the LS estimate obtained from the sample; it is the population parameter and the target of estimation.
  - The N-vectors $\boldsymbol{\xi}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are *not* orthogonal to the predictor columns $\boldsymbol{X}_j$ in the sample; in general $\langle \boldsymbol{\xi}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\epsilon}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\eta}, \boldsymbol{X}_j \rangle \neq 0$, even though the associated random variables are orthogonal to X_j in the population: $\boldsymbol{E}[\,\xi X_j] = \boldsymbol{E}[\,\epsilon X_j] = \boldsymbol{E}[\,\eta(\vec{\boldsymbol{X}}) X_j] = \;0$.

3.3 Quantities associated with LS estimation

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

$$\hat{\boldsymbol{\beta}} \;=\; (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^T \;=\; \text{argmin}_{\boldsymbol{\beta}} \, \|\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2 \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,. \tag{20}$$

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

Hat or projection matrix:	$\boldsymbol{H} \;=\; \boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$	(21)
LS fits:	$\hat{\boldsymbol{Y}} \;=\; \boldsymbol{X}\hat{\boldsymbol{\beta}} \;=\; \boldsymbol{H}\boldsymbol{Y}$	(22)
Residuals:	$\boldsymbol{r} \;=\; \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$	(23)

  [Note: Sample residual vector $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \neq$ population residual vector $\boldsymbol{\xi} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$.]

- **Adjustment formulas for samples**: Again, we introduce adjusted variables to express a multiple regression as a collection of simple regressions. The sample version of adjustment is immediately useful in suggesting diagnostic plots, whereas the population version above served to provide targets of estimation. In a sample, we collect all predictor columns other than $\boldsymbol{X}_j$ in a $N \times p$ matrix $\boldsymbol{X}_{\setminus j} = (\boldsymbol{1}, ..., \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, ...)$, and we define

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

  The sample-adjusted predictor is

$$\boldsymbol{X}_{j\bullet} \;=\; \boldsymbol{X}_j - \hat{\boldsymbol{\beta}}_{j\bullet}^T \boldsymbol{X}_{\setminus j} \;=\; (\boldsymbol{I} - \boldsymbol{H}_{\setminus j}) \boldsymbol{X}_j \, . \tag{24}$$

  where $\boldsymbol{H}_{\setminus j} = \boldsymbol{X}_{\setminus j}(\boldsymbol{X}_{\setminus j}^T \boldsymbol{X}_{\setminus j})^{-1} \boldsymbol{X}_{\setminus j}^T$ 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 $\boldsymbol{X}_{j\bullet}$:

$$\hat{\beta}_j \;=\; \frac{\langle \boldsymbol{Y}, \boldsymbol{X}_{j\bullet} \rangle}{\|\boldsymbol{X}_{j\bullet}\|^2} \, . \tag{25}$$

  [An important distinction is that the vector $\boldsymbol{X}_{j\bullet}$ is sample-adjusted, whereas the random variable $X_{j\bullet}$ is population-adjusted. This makes the notation "$X_{i,j\bullet}$" ambiguous: it could refer to the i'th component of the sample-adjusted vector $\boldsymbol{X}_{j\bullet}$ or to the i'th case drawn from the population-adjusted random variable $X_{j\bullet}$. These are not the same. We will therefore avoid using "$X_{i,j\bullet}$", which is possible without difficulty.]

4 Decomposition of the LS Estimate according to its Sources of Variation

When the predictors are random and the linear model is treated as an approximation rather than a generative truth, the sampling variation of the LS estimate $\hat{\boldsymbol{\beta}}$ can be additively decomposed into two components: one component due to error, and another component due to nonlinearity interacting with randomness of the predictors. This decomposition is a direct reflection of the decomposition of the population residual into error and nonlinearity, $\boldsymbol{\xi} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$, according to (11). In this section we state asymptotic normality for each part of the decomposition. The relevance of the decomposition is that it explains what the paired bootstrap estimates, while the associated asymptotic normalities are necessary to justify the paired bootstrap.

In the conventional analysis of linear models, 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 allowed, the target of estimation is the population LS solution $\boldsymbol{\beta}$ as defined in (4)-(7). In this case, $\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$ is a random variable that sits between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; (\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \;+\; (\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta}) \tag{26}$$

This decomposition corresponds to the decomposition $\boldsymbol{\xi} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$ as the following lemma shows. The lemma gives both vectorized and componentwise versions, as the former provide easy proofs, while the latter will result in simple diagnostics.

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

they will be prefixed as follows:

$$\textit{Total EO:} \quad \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\xi}, \qquad \hat{\beta}_j - \beta_j = \frac{\langle \boldsymbol{X}_{j\bullet}, \boldsymbol{\xi}\rangle}{\|\boldsymbol{X}_{j\bullet}\|^2}. \tag{27}$$

$$\textit{Error EO:} \quad \hat{\boldsymbol{\beta}} - \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}, \qquad \hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] = \frac{\langle \boldsymbol{X}_{j\bullet}, \boldsymbol{\epsilon}\rangle}{\|\boldsymbol{X}_{j\bullet}\|^2}. \tag{28}$$

$$\textit{Nonlinearity EO:} \quad \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}, \qquad \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j = \frac{\langle \boldsymbol{X}_{j\bullet}, \boldsymbol{\eta}\rangle}{\|\boldsymbol{X}_{j\bullet}\|^2}. \tag{29}$$

Proof: It is sufficient to consider the vectorized versions. The lemma is a consequence of similar equations for $\boldsymbol{Y}$, $\boldsymbol{\mu}$ and $\boldsymbol{X}\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{30}$$

$$\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\mu} \tag{31}$$

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}) \tag{32}$$

Of these equations the first is the definition of $\hat{\boldsymbol{\beta}}$, the middle follows from the first using $\boldsymbol{E}[\,Y_i|\vec{\boldsymbol{X}}_i] = \mu(\vec{\boldsymbol{X}}_i)$, and the last is a tautology. The equations of the lemma follow by forming the pairwise differences (30)-(32), (30)-(31), and (31)-(32), respectively, and recalling the vectorized definitions (19): $\boldsymbol{\xi} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}$. □

The EOs are not actionable quantities in real data analysis, but they will be useful in comparing unconditional (bootstrap) and conditional (linear) inference.

Proposition: *The three EOs follow central limit theorems:*

$$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\boldsymbol{0},\ \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\right) \tag{33}$$

$$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\boldsymbol{0},\ \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\right) \tag{34}$$

$$N^{1/2}\,(\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\boldsymbol{0},\ \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\right) \tag{35}$$

Proof: It is sufficient to prove the first statement, say, as the others follow the same way.

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

We used $\boldsymbol{E}[\,\xi\,] = 0\,(= \boldsymbol{E}[\,\epsilon\,] = \boldsymbol{E}[\,\eta\,])$ according to (15). It follows that

$$\boldsymbol{V}[\,\xi\vec{\boldsymbol{X}}] = \boldsymbol{E}[\,\xi^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] = \boldsymbol{E}[\,\boldsymbol{E}[\,\xi^2|\vec{\boldsymbol{X}}]\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] = \boldsymbol{E}[\,\rho^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] \qquad \square$$

The asymptotic variance/covariance matrices in the proposition have the well-known sandwich or butterfly form. Asymptotic normality can also be expressed for each $\hat{\beta}_j$ separately using population adjustment:

Corollary: *Each component of an estimation offset follows a central limit theorem.*

$$N^{1/2}(\hat{\beta}_j - \beta_j) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(0, \frac{\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}}) {X_{j\bullet}}^2]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}\right) \tag{36}$$

$$N^{1/2}(\hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}]) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(0, \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}}) {X_{j\bullet}}^2]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}\right) \tag{37}$$

$$N^{1/2}(\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}] - \beta_j) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(0, \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}}) {X_{j\bullet}}^2]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}\right) \tag{38}$$

Proof: By the proposition above we know that each component is marginally asymptotically normal. Hence we only need to verify the form of the asymptotic variance. This in turn is achieved by specializing the asymptotic variance/covariance matrix of the proposition above to the single predictor case with $X_{j\bullet}$ as the random variable underlying the predictor. □

Corollary: *The unconditional or marginal bias of $\hat{\beta}_j$ vanishes faster than $N^{-1/2}$:*

$$N^{1/2}\,(\boldsymbol{E}[\hat{\beta}_j] - \beta_j) \;\longrightarrow\; 0 \quad (N \longrightarrow \infty). \tag{39}$$

Proof: This follows from (36). □

[xxxxxxxxx Preliminary note – might go elsewhere:
In econometrics it has long been exploitet that (33) suggests a standard error estimate — the so-called "sandwich estimator" — that is asymptotically correct even when the linear model is incorrect. It is obtained by estimating $\rho^2(\vec{\boldsymbol{X}})$ with $(Y - \hat{\boldsymbol{\beta}}^T \vec{\boldsymbol{X}})^2$, which is asymptotically unbiased: $\boldsymbol{E}[(Y - \hat{\boldsymbol{\beta}}^T \vec{\boldsymbol{X}})^2|\vec{\boldsymbol{X}}] \to \rho^2(\vec{\boldsymbol{X}})$. The sandwich estimator has come under criticism for being inefficient (Kauermann and Carroll 2001). This inefficiency problem might get corrected by drawing on the adjustment version (36) and noting that

$$\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}}) {X_{j\bullet}}^2] \;=\; \boldsymbol{E}[\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}}) \,|\, {X_{j\bullet}}^2]\, {X_{j\bullet}}^2]\,.$$

The expression $\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}}) \,|\, {X_{j\bullet}}^2]$ can be estimated nonparametrically in the sample by smoothing $\boldsymbol{r}^2 = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2$ (= the vector of squared residuals) against ${\boldsymbol{X}_{j\bullet}}^2$ (= the vector of squared adjusted predictor values). Such smoothing should remove some of the noisiness and hence inefficiency of the sandwich estimator.]

5 Standard Errors of Regression Coefficients

We compare conventional standard errors with the actual (marginal, unconditional) standard errors. The latter correctly include the additional variation due to the nonlinearity and randomness of $\boldsymbol{X}$. In what follows we distinguish between (squared) *standard error* $\boldsymbol{V}[\hat{\beta}_j]$, *conditional standard error* $\boldsymbol{V}[\hat{\beta}_j|\boldsymbol{X}]$, and *standard error estimate* $\hat{\boldsymbol{V}}[\hat{\beta}_j]$.

1. The (squared) *conventional standard error* of $\hat{\beta}_j$ is estimated by $\hat{\sigma}^2/\|\boldsymbol{X}_{j\bullet}\|^2$ where one uses the usual estimate of σ^2 from the linear model as if the model were correct: $\hat{\sigma}^2 = \frac{1}{N-p-1}\,\|\boldsymbol{r}\|^2$. We calculate its conditional mean allowing for both nonlinearity and heteroskedasticity (see Appendix B.1):

$$\boldsymbol{E}[\hat{\sigma}^2|\boldsymbol{X}] \;=\; \tfrac{1}{N-p-1}\left(\mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{D}_{\boldsymbol{\sigma}^2}) + \|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\eta}\|^2\right), \tag{40}$$

where $\boldsymbol{D}_{\boldsymbol{\sigma}^2}$ is the diagonal matrix with the values $\sigma^2(\vec{\boldsymbol{X}}_i)$ in its diagonal. Thus the conventional standard error estimate $\hat{\sigma}^2/\|\boldsymbol{X}_{j\bullet}\|^2$ is $\boldsymbol{X}$-conditionally an unbiased estimate of the following:

$$\mathrm{SE}^2_{\mathrm{conv}}[\hat{\beta}_j|\boldsymbol{X}] \;:=\; \boldsymbol{E}\left[\frac{\hat{\sigma}^2}{\|\boldsymbol{X}_{j\bullet}\|^2}\middle|\boldsymbol{X}\right] \;=\; \frac{\frac{1}{N-p-1}\mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{D}_{\boldsymbol{\sigma}^2})}{\|\boldsymbol{X}_{j\bullet}\|^2} \;+\; \frac{\frac{1}{N-p-1}\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\eta}\|^2}{\|\boldsymbol{X}_{j\bullet}\|^2} \tag{41}$$

On the right hand side, the first term accounts for error (possibly heteroscedastic), the second term for nonlinearity. Under homoscedasticity ($\boldsymbol{D}_{\boldsymbol{\sigma}^2} = \sigma^2\boldsymbol{I}$) and linearity ($\boldsymbol{\eta} = \boldsymbol{0}$) it is true that $\mathrm{SE}^2_{\mathrm{conv}}[\hat{\beta}_j|\boldsymbol{X}] = \boldsymbol{V}[\hat{\beta}_j|\boldsymbol{X}]$, hence this is an appropriate conditional squared standard error for $\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}]$, which under linearity equals β_j. In the presence of nonlinearity and/or heteroscedasticity, however, $\mathrm{SE}^2_{\mathrm{conv}}[\hat{\beta}_j|\boldsymbol{X}]$ is not a correct squared standard error for β_j, not even conditionally for $\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}]$, as the next item will show.

2. The (squared) unconditional or *marginal standard error* is $\boldsymbol{V}[\hat{\beta}_j]$. Its theoretical relevance is that it is the true standard error that accounts for sampling variability due to error *and* due to randomness of the predictors. Its practical importance is that it can and should be estimated by sandwich estimators or the paired bootstrap. It can be decomposed as follows:

$$\mathrm{SE}^2_{\mathrm{marg}}[\hat{\beta}_j] \;:=\; \boldsymbol{V}\left[\hat{\beta}_j\right] \;=\; \boldsymbol{E}\left[\boldsymbol{V}[\hat{\beta}_j|\boldsymbol{X}]\right] \;+\; \boldsymbol{V}\left[\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}]\right] \tag{42}$$

$$\;=\; \boldsymbol{E}\left[\frac{\boldsymbol{X}_{j\bullet}{}^T\boldsymbol{D}_{\boldsymbol{\sigma}^2}\boldsymbol{X}_{j\bullet}}{\|\boldsymbol{X}_{j\bullet}\|^4}\right] \;+\; \boldsymbol{V}\left[\frac{\langle\boldsymbol{X}_{j\bullet},\boldsymbol{\eta}\rangle}{\|\boldsymbol{X}_{j\bullet}\|^2}\right]. \tag{43}$$

In the last equality we drew on (25).

Thus both the conventional and the marginal standard error can be additively decomposed into contributions due to $\sigma^2(\vec{\boldsymbol{X}})$ (error) and $\eta^2(\vec{\boldsymbol{X}})$ (nonlinearity). For a simplified asymptotic analysis, we form the scaled limits:

Proposition and Definition: *Defining* $\mathrm{SE}^2_{\mathrm{conv}}[\hat{\beta}_j] := \boldsymbol{E}[\mathrm{SE}^2_{\mathrm{conv}}[\hat{\beta}_j|\boldsymbol{X}]]$ *and* $\mathrm{SE}^2_{\mathrm{marg}}[\hat{\beta}_j]$ *as in* (42)*, we have:*

$$N\,\mathrm{SE}^2_{\mathrm{conv}}(\hat{\beta}_j) \;\overset{N\to\infty}{\longrightarrow}\; \underbrace{\frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[{X_{j\bullet}}^2]}}_{\boldsymbol{AV}^{(j)}_{\mathrm{conv}}(\sigma^2(\vec{\boldsymbol{X}}))} + \underbrace{\frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[{X_{j\bullet}}^2]}}_{\boldsymbol{AV}^{(j)}_{\mathrm{conv}}(\eta^2(\vec{\boldsymbol{X}}))} = \underbrace{\frac{\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[{X_{j\bullet}}^2]}}_{\boldsymbol{AV}^{(j)}_{\mathrm{conv}}(\rho^2(\vec{\boldsymbol{X}}))} \tag{44}$$

$$N\,\mathrm{SE}^2_{\mathrm{marg}}(\hat{\beta}_j) \;\overset{N\to\infty}{\longrightarrow}\; \underbrace{\frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}}_{\boldsymbol{AV}^{(j)}_{\mathrm{marg}}(\sigma^2(\vec{\boldsymbol{X}}))} + \underbrace{\frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}}_{\boldsymbol{AV}^{(j)}_{\mathrm{marg}}(\eta^2(\vec{\boldsymbol{X}}))} = \underbrace{\frac{\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}}_{\boldsymbol{AV}^{(j)}_{\mathrm{marg}}(\rho^2(\vec{\boldsymbol{X}}))} \tag{45}$$

The proofs of (44) and (45) are obtained from the sum representations in (41) and (43). For details of (44) see Appendix B.3; essentially $\boldsymbol{H}$ and $p+1$ in (41) become negligible for p fixed as $N \to \infty$. The limits in (45) are the asymptotic variances (36)-(38) of the preceding section.

Structurally all three terms $\boldsymbol{AV}^{(j)}_{\mathrm{conv}}(...)$ have the same form, and so do the three terms $\boldsymbol{AV}^{(j)}_{\mathrm{marg}}(...)$, justifying the notation. Only $\boldsymbol{AV}^{(j)}_{\mathrm{marg}}(...)$ represent inferentially valid asymptotic variances, while $\boldsymbol{AV}^{(j)}_{\mathrm{conv}}(...)$ are generally valid only when $\sigma^2(\vec{\boldsymbol{X}}) \equiv \sigma^2$ and $\eta \equiv 0$.

6 Asymptotic Comparison of Conventional and Marginal Standard Errors

We show 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 $\rho^2(\vec{\boldsymbol{X}})$. To this end we form the ratios $\boldsymbol{RAV}_j(...)$ as follows:

Definition and Lemma: *Ratios of marginal and conventional Asymptotic Variances*

$$\boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}})) \;:=\; \frac{\boldsymbol{AV}^{(j)}_{\text{marg}}(\sigma^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{\text{conv}}(\sigma^2(\vec{\boldsymbol{X}}))} \;=\; \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})X_j{\bullet}^2]}{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_j{\bullet}^2]} \tag{46}$$

$$\boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}})) \;:=\; \frac{\boldsymbol{AV}^{(j)}_{\text{marg}}(\eta^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{\text{conv}}(\eta^2(\vec{\boldsymbol{X}}))} \;=\; \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})X_j{\bullet}^2]}{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_j{\bullet}^2]} \tag{47}$$

$$\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}})) \;:=\; \frac{\boldsymbol{AV}^{(j)}_{\text{marg}}(\rho^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{\text{conv}}(\rho^2(\vec{\boldsymbol{X}}))} \;=\; \frac{\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}})X_j{\bullet}^2]}{\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_j{\bullet}^2]} \tag{48}$$

The second equality on each line follows from (44) and (45). The ratios express by how much the conventional asymptotic variances need to multiplied to match the proper marginal asymptotic variances. Among the three ratios the relevant one for the overall comparison of conventional and marginal inference is $\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}}))$. For example, if $\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}})) = 4$, 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 $\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}})) = 1$, the conventional standard error for $\hat{\beta}_j$ is asymptotically correct;
- if $\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}})) > 1$, the conventional standard error for β_j is asymptotically too small/optimistic;
- if $\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}})) < 1$, 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 heteroscedasticity and/or nonlinearity contribute asymptotically to the defects of conventional standard errors. Structurally, the three ratios are inner products between the normalized squared quantities

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

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. To analyze the range of $\boldsymbol{RAV}_j$ and to take advantage of the identical form of $\boldsymbol{RAV}_j$ in all three cases, we will write $f(\vec{\boldsymbol{X}})$ for any of $\sigma(\vec{\boldsymbol{X}})$, $\eta(\vec{\boldsymbol{X}})$ and $\rho(\vec{\boldsymbol{X}})$, and analyze the range of $\boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}}))$ over scenarios $f(\vec{\boldsymbol{X}})$. It is easy to see that generally

$$\sup_f \boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) = \infty\,, \qquad \inf_f \boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) = 0\,, \tag{49}$$

where $f(\vec{\boldsymbol{X}})$ ranges over functions with suitable moments such that $\boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}}))$ exists. The limits (49) become fairly obvious from the following equivalent form of $\boldsymbol{RAV}_j$:

$$\boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) \;=\; \frac{\boldsymbol{E}[\boldsymbol{E}[f^2(\vec{\boldsymbol{X}})\,|\,X_j{\bullet}^2]\,X_j{\bullet}^2]}{\boldsymbol{E}[f^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_j{\bullet}^2]} \tag{50}$$

Thus $\boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}}))$ attains

- a large value if $\boldsymbol{E}[f^2(\vec{\boldsymbol{X}}) \,|\, {X_j}_\bullet{}^2]$ and ${X_j}_\bullet{}^2$ are positively associated, which is the case for example if $\boldsymbol{E}[f^2(\vec{\boldsymbol{X}}) \,|\, {X_j}_\bullet{}^2]$ is a monotone increasing function of ${X_j}_\bullet{}^2$;
- a value near zero if $\boldsymbol{E}[f^2(\vec{\boldsymbol{X}}) \,|\, {X_j}_\bullet{}^2]$ and ${X_j}_\bullet{}^2$ are negatively associated, which is the case for example if $\boldsymbol{E}[f^2(\vec{\boldsymbol{X}}) \,|\, {X_j}_\bullet{}^2]$ is a monotone decreasing function of ${X_j}_\bullet{}^2$.

We will illustrate the above with some simple scenario constructions.

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:*

$$\eta_t(X) \;=\; \frac{1_{[|X|>t]} - p(t)}{\sqrt{p(t)(1-p(t))}}\,, \qquad \textit{where} \quad p(t) = \boldsymbol{P}[|X| > t]\,, \tag{51}$$

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:* $\boldsymbol{P}[X = 0] = 0$.

(See Appendix B.5 for the simple proof). 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 heteroscedastic 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 heteroscedastic error variances as follows:*

$$\sigma_t^2(X) \;=\; \frac{(1_{[|X|>t]} - p(t))^2}{p(t)(1-p(t))}\,, \qquad \textit{where} \quad p(t) = \boldsymbol{P}[|X| > t]\,, \tag{52}$$

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:* $\boldsymbol{P}[X = 0] = 0$.

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.

In the end, of course, it is $\rho^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ that determines the gap between conventional and marginal inferences. The following lemma can be used to examine when $\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}}))$ approaches ∞ or 0:

Lemma: *Define weights*

$$w_\sigma = \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}})]}\,, \qquad w_\eta = \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\rho^2(\vec{\boldsymbol{X}})]}\,, \tag{53}$$

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

$$\boldsymbol{RAV}_j(\rho^2(\vec{\boldsymbol{X}})) \;=\; w_\sigma\, \boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}})) + w_\eta\, \boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}}))\,. \tag{54}$$

Corollary: *If* $\rho_t^2(\vec{\boldsymbol{X}}) = \sigma_t^2(\vec{\boldsymbol{X}}) + \eta_t^2(\vec{\boldsymbol{X}})$ *is a one-parameter family of scenarios, and if* $w_{\sigma_t} = w_\sigma > 0$ *and* $w_{\eta_t} = w_\eta > 0$ *from* (53) *are fixed and strictly positive, then we have*

$\lim_t \boldsymbol{RAV}_j(\rho_t^2(\vec{\boldsymbol{X}})) = \infty$ *if* $\lim_t \boldsymbol{RAV}_j(\sigma_t^2(\vec{\boldsymbol{X}})) = \infty$ ***or*** $\lim_t \boldsymbol{RAV}_j(\eta_t^2(\vec{\boldsymbol{X}})) = \infty$*;*

$\lim_t \boldsymbol{RAV}_j(\rho_t^2(\vec{\boldsymbol{X}})) = 0$ *if both* $\lim_t \boldsymbol{RAV}_j(\sigma_t^2(\vec{\boldsymbol{X}})) = 0$ ***and*** $\lim_t \boldsymbol{RAV}_j(\eta_t^2(\vec{\boldsymbol{X}})) = 0$.

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

7 Submodels

Following the notation in the PoSI article, we consider submodels given by indices $\mathrm{M} = \{j_1, ..., j_m\}$ with predictor submatrices $\boldsymbol{X}_\mathrm{M} = (\boldsymbol{X}_{j_1}, ..., \boldsymbol{X}_{j_m})$ and associated vectors of parameters and their LS estimates $\boldsymbol{\beta}_\mathrm{M}$ and $\hat{\boldsymbol{\beta}}_\mathrm{M}$, respectively. We know the asymptotic normality of $\hat{\boldsymbol{\beta}}_\mathrm{M}$ for each M, but we are now also interested in the joint asymptotic normality of $(\hat{\boldsymbol{\beta}}_\mathrm{M}^T, \hat{\boldsymbol{\beta}}_{\mathrm{M}'}^T)^T$ for two submodels M and M′. While the asymptotic variance for $\hat{\boldsymbol{\beta}}_\mathrm{M}$ is

$$\boldsymbol{AV}(\hat{\boldsymbol{\beta}}_\mathrm{M}) \;=\; \boldsymbol{E}[\vec{\boldsymbol{X}}_\mathrm{M}\vec{\boldsymbol{X}}_\mathrm{M}^T]^{-1}\, \boldsymbol{E}[\xi_\mathrm{M}^2\vec{\boldsymbol{X}}_\mathrm{M}\vec{\boldsymbol{X}}_\mathrm{M}^T]\, \boldsymbol{E}[\vec{\boldsymbol{X}}_\mathrm{M}\vec{\boldsymbol{X}}_\mathrm{M}^T]^{-1},$$

using the submodel-specific "population residual" variable $\xi_\mathrm{M} = Y - \vec{\boldsymbol{X}}_\mathrm{M}^T\boldsymbol{\beta}_\mathrm{M}$, the cross-asymptotic covariance between $\hat{\boldsymbol{\beta}}_\mathrm{M}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{M}'}$ is

$$\boldsymbol{AV}(\hat{\boldsymbol{\beta}}_\mathrm{M}, \hat{\boldsymbol{\beta}}_{\mathrm{M}'}) \;=\; \boldsymbol{E}[\vec{\boldsymbol{X}}_\mathrm{M}\vec{\boldsymbol{X}}_\mathrm{M}^T]^{-1}\, \boldsymbol{E}[\xi_\mathrm{M}\xi_{\mathrm{M}'}\vec{\boldsymbol{X}}_\mathrm{M}\vec{\boldsymbol{X}}_{\mathrm{M}'}^T]\, \boldsymbol{E}[\vec{\boldsymbol{X}}_{\mathrm{M}'}\vec{\boldsymbol{X}}_{\mathrm{M}'}^T]^{-1}.$$

A The Meaning of Regression Slopes as Averages of Case-Based Slopes

In this section we give simple interpretations to slopes when the actual response function is not linear in the predictors. We first condition on the design and then consider the random design case. It will also be sufficient to consider simple linear regression only, keeping in mind that all formulas apply to multiple regression by replacing the single predictor with the j'th predictor adjusted for all other predictors.

For a sample (Y_i, X_i) $(i = 1, ..., N)$ of single-predictor/response data, the LS slope estimate can be represented as follows:

$$\hat{\beta} \;=\; \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_k (X_k - \bar{X})^2} \;=\; \frac{\sum_i \frac{Y_i - \bar{Y}}{X_i - \bar{X}} (X_i - \bar{X})^2}{\sum_k (X_k - \bar{X})^2} \;=\; \sum_i w_i \, b_i \,, \tag{55}$$

where

$$b_i \;=\; \frac{Y_i - \bar{Y}}{X_i - \bar{X}} \qquad \text{and} \qquad w_i \;=\; \frac{(X_i - \bar{X})^2}{\sum_k (X_k - \bar{X})^2}$$

are case-based slopes and weights. The slope for the i'th case is based on the line segment that reaches from the center point $(\bar{X}, \bar{Y})$ to the data point (X_i, Y_i). The weights grow with the square of the distance of X_i from $\bar{X}$. This is consistent with the variance of the case-based slopes b_i: $\boldsymbol{V}[b_i|\boldsymbol{X}] \;=\; \frac{n-1}{n} \, \sigma^2/(X_i - \bar{X})^2$, assuming homoskedasticity only, $\boldsymbol{V}[Y_i|X_i] = \sigma^2$, but not linearity of $\mu(X_i) = \boldsymbol{E}[Y_i|X_i]$. Plausibly, $w_i \sim 1/\boldsymbol{V}[b_i] \sim (X_i - \bar{X})^2$, and $\hat{\beta}$ is therefore the LS solution of a weighted mean problem:

$$\hat{\beta} \;=\; \text{argmin}_{\tilde{\beta}} \sum_i w_i \, (b_i - \tilde{\beta})^2,$$

which is the same as the unweighted LS problem

$$\hat{\beta} \;=\; \text{argmin}_{\tilde{\beta}} \sum_i ((Y_i - \bar{Y}) - \tilde{\beta} \, (X_i - \bar{X}))^2.$$

[A well-known variant, used for example in Gelman and Park (2008), represents the slope estimate as follows:

$$\hat{\beta} \;=\; \sum_{i,i'} w_{i,i'} \, b_{i,i'} \,, \qquad \text{where} \qquad w_{i,i'} = \frac{(X_i - X_{i'})^2}{\sum_{k,k'} (X_k - X_{k'})^2} \qquad \text{and} \qquad b_{i,i'} = \frac{Y_i - Y_{i'}}{X_i - X_{i'}}$$

are pairwise weights and slopes.]

We can carry the preceding over to $\boldsymbol{E}[\hat{\beta}|\boldsymbol{X}]$, which is the target of LS estimation conditional on the design $\boldsymbol{X}$. We make no assumption of linearity about $\mu_i = \boldsymbol{E}[Y_i|\boldsymbol{X}]$, which can in fact be entirely arbitrary. Taking the conditional expectation in (55) given $\boldsymbol{X}$, we obtain

$$\boldsymbol{E}[\hat{\beta}|\boldsymbol{X}] \;=\; \frac{\sum_i (\mu_i - \bar{\mu}) \cdot (X_i - \bar{X})}{\sum_k (X_k - \bar{X})^2} \;=\; \sum_i w_i \, \beta_i \,, \tag{56}$$

where $\beta_i = (\mu_i - \bar{\mu})/(X_i - \bar{X})$. Thus $\boldsymbol{E}[\hat{\beta}|\boldsymbol{X}]$, too, is an average of case-based slopes, for line segments from $(\bar{X}, \bar{\mu})$ to (X_i, μ_i). [This is illustrated in Figure 3.]

Finally, we can extend the above to the population when the design $\boldsymbol{X}$ is considered random, no conditioning is applied, and again no assumption of linearity about $\mu(X) = \boldsymbol{E}[Y|X]$ is made:

$$\begin{aligned} \beta \;&=\; \text{argmin}_{\tilde{\beta}} \, \boldsymbol{E}\left[\Big((Y - \boldsymbol{E}[Y]) - \tilde{\beta}(X - \boldsymbol{E}[X])\Big)^2\right] \\ &=\; \frac{\boldsymbol{E}\left[(\mu(X) - \boldsymbol{E}[\mu(X)]) \, (X - \boldsymbol{E}[X])\right]}{\boldsymbol{E}[(X - \boldsymbol{E}[X])^2]} \\ &=\; \boldsymbol{E}[w(X) \, \beta(X)] \,, \end{aligned}$$

where

$$\beta(X) = \frac{\mu(X) - \boldsymbol{E}[\mu(X)]}{X - \boldsymbol{E}[X]} \quad \text{and} \quad w(X) = \frac{(X - \boldsymbol{E}[X])^2}{\boldsymbol{E}[(X - \boldsymbol{E}[X])^2]},$$

and, of course: $\boldsymbol{E}[\mu(X)] = \boldsymbol{E}[Y]$.

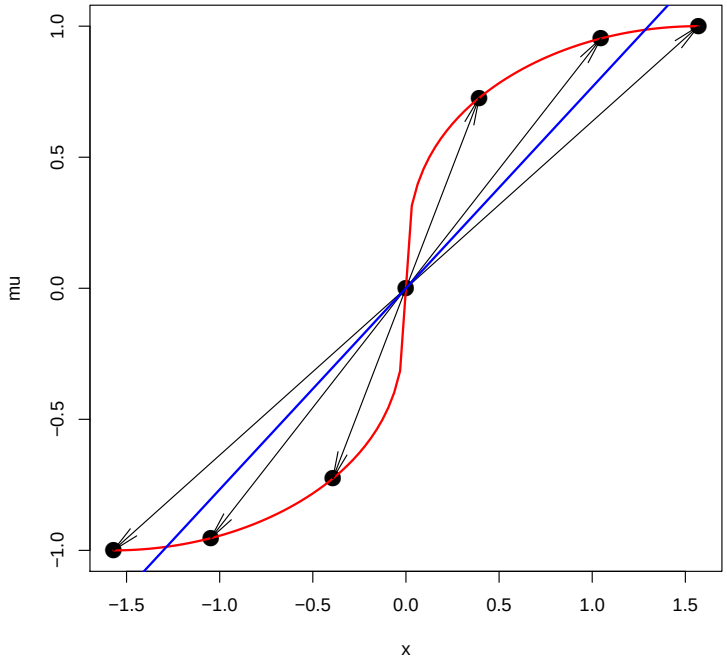

Figure 3: *The dots show points* (X_i, μ_i) *that do not fall on a line; they fall on the red curve. The point* $(0, 0)$ *is the center* $(\bar{X}, \bar{\mu})$*. The arrows show the segments from the center to the points* (X_i, μ_i)*; their slopes are* β_i*. The average slope* β *is shown in blue. It is relatively flat because the weights are large for outlying values* X_i*.*

B Proofs

B.1 Conditional Expectation of RSS

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

$$\begin{aligned}
\boldsymbol{E}[\|\boldsymbol{r}\|^2|\boldsymbol{X}] &= E[\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}|\boldsymbol{X}] && (57)\\
&= \boldsymbol{E}[(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon})'(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon})|\boldsymbol{X}] && (58)\\
&= \boldsymbol{E}[(\boldsymbol{\eta} + \boldsymbol{\epsilon})^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{\eta} + \boldsymbol{\epsilon})|\boldsymbol{X}] && (59)\\
&= \mathrm{tr}(\boldsymbol{E}[(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\eta} + \boldsymbol{\epsilon})^T|\boldsymbol{X}]) && (60)\\
&= \mathrm{tr}((\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{\eta}\boldsymbol{\eta}^T + \boldsymbol{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]) && (61)\\
&= \mathrm{tr}((\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{\eta}\boldsymbol{\eta}^T + \boldsymbol{D}_{\boldsymbol{\sigma}^2}) && (62)\\
&= |(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\eta}|^2 + \mathrm{tr}((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{D}_{\boldsymbol{\sigma}^2}) && (63)
\end{aligned}$$

B.2 Limit of Squared Adjusted Predictors

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

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

B.3 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 $\|\boldsymbol{H}\boldsymbol{\eta}\|^2$:

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

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

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

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

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

$$N\,\frac{\frac{1}{N-p-1}\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\eta}\|^2}{\|\boldsymbol{X}_{j\bullet}\|^2} \xrightarrow{\boldsymbol{P}} \frac{\boldsymbol{E}[\eta(\vec{\boldsymbol{X}})^2]}{\boldsymbol{E}[{X_{j\bullet}}^2]}. \tag{64}$$

B.4 Asymptotic Normality in Terms of Adjustment

We gave the asymptotic limit of the conditional bias in vectorized form after (33)-(35). Here we derive the equivalent element-wise limit using adjustment to show (36)-(38). The variance of the conditional bias is the marginal inflator of SE.

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

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

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

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

B.5 A Family of Nonlinearities with Extreme *RAV*

We refer to the example constructed in Section 6 and recall that both $\boldsymbol{E}[X^2] = 1$ and $\boldsymbol{E}[\eta_t(X)^2] = 1$. We abbreviate $\bar{p}(t) = 1 - p(t)$.

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

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

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

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

References

[1] 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.

[2] Belseley, 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.

[3] 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.) xxx–xxx, Place: Publisher.

[4] 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.

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

[6] 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.

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

[8] 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.

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

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

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

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

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

[14] Kauermann, G. and Carroll, R. J. (2001). *A note on the efciency of sandwich covariance matrix estimation*, *Journal of the American Statistical Association* **96**(456), 1387-1396.

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