Journal club – Models as Approximations

Anders Munch

June 13, 2022

The paper

Models as Approximations I: Consequences Illustrated with Linear Regression

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

Wharton – University of Pennsylvania[‡] and Citadel[§] and UNC at Chapel Hill[¶]

Abstract.

In the early 1980s Halbert White inaugurated a "model-robust" form of statistical inference based on the "sandwich estimator" of standard error. This estimator is known to be "heteroskedasticity-consistent", but it is less well-known to be "nonlinearity-consistent" as well. Nonlinearity, however, raises fundamental issues because in its presence regressors are not ancillary, hence can't be treated as fixed. The consequences are deep: (1) population slopes need to be re-interpreted as statistical functionals obtained from OLS fits to largely arbitrary joint x-y distributions; (2) the meaning of slope parameters needs to be rethought; (3) the regressor distribution affects the slope parameters; (4) randomness of the regressors becomes a source of sampling variabil-

Discussion paper

- Memorial Issue of *Statitical Science* for Lawrence D. Brown
- Paper in two parts: Special case of linear regression (part I, Buja et al. [2019a]) and general case (part II, Buja et al. [2019b]). Part II defines a notion of "well-specification for regression functionals" and propose diagnostic tools.

Comments by

- Ghanem and Kuffner (UC Davis and Washington University in St. Louis)
- Rinaldo, Tibshirani, and Wasserman (Carnegie Mellon University, Pittsburgh)
- Whitney, Shojaie, and Carone (Imperial College London, University of Washington, Seattle)
- and others.

Overall idea

Understand what is estimated with linear a linear regression (or more general M-estimators) when the model is mis-specified.

Define the parameter β as a functional defined on the data distribution $P \in \mathcal{P}$:

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

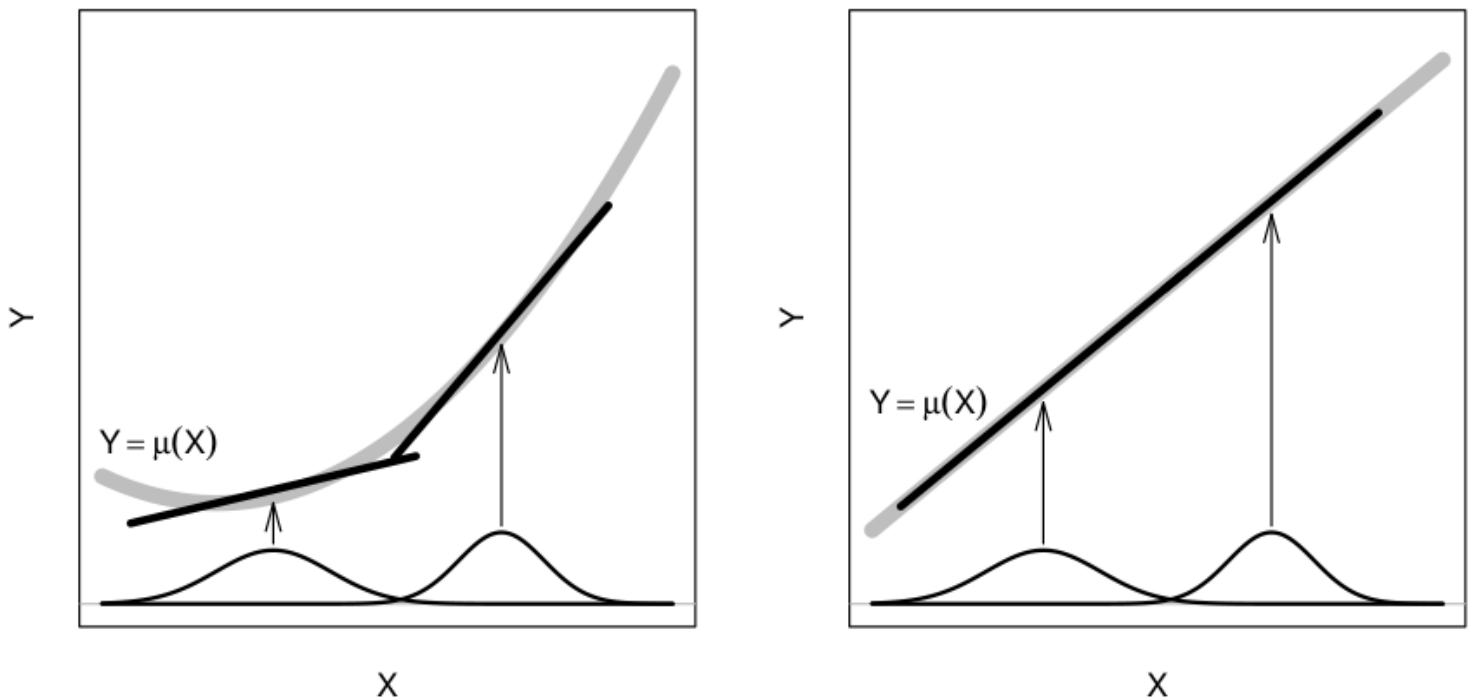
The main definition concerns ***the best population linear approximation*** to Y, which is the linear function $l(\vec{\boldsymbol{X}}) = \boldsymbol{\beta}'\,\vec{\boldsymbol{X}}$ with coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ given by

$$\begin{aligned} \boldsymbol{\beta}(\boldsymbol{P}) \;&\stackrel{\Delta}{=} \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}\, \boldsymbol{E}[(Y - \boldsymbol{\beta}'\,\vec{\boldsymbol{X}})^2] \qquad = \boldsymbol{E}[\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\,Y\,] \\ &= \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}\, \boldsymbol{E}[(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\,\vec{\boldsymbol{X}})^2] \;= \boldsymbol{E}[\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}\,\mu(\vec{\boldsymbol{X}})]. \end{aligned}$$

One consequence

4.2 Implications of the Dependence of Slopes on Regressor Distributions

A first practical implication, illustrated by Figure 2, is that two empirical studies that use the same regressors, the same response, and the same model, may yet estimate different parameter values, $\boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2)$. This possibility arises even if the true response surface $\mu(\vec{\boldsymbol{x}})$ is identical between the studies. The reason is model misspecification and differences between the regressor distributions in the two studies. Here is therefore a potential cause of so-called "parameter hetero-



Is this a problem?

This seems to only be a problem because we do not understand how to interpret $\beta(P)$.

Average treatment effect (ATE)

The ATE,

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}_P\left[\mathbb{E}_P[Y \mid X, A = 1] - \mathbb{E}_P[Y \mid X, A = 0]\right],$$

should naturally depend on the background distribution of X.

Conditional average treatment effect (CATE)

The CATE,

$$\mathbb{E}[Y^1 - Y^0 \mid X = x] = \mathbb{E}_P[Y \mid X = x, A = 1] - \mathbb{E}_P[Y \mid X = x, A = 0],$$

would not depend on the background distribution X (right?).

Interpretation of "slopes" in the presence of non-linearity

Population parameters β can be represented as weighted averages of ...

- **case-wise slopes**: For a random case (x, y) we have

$$\beta = \boldsymbol{E}[\, w\, b\,], \quad \text{where} \quad b \triangleq \frac{y}{x}, \qquad w \triangleq \frac{x^2}{\boldsymbol{E}[\, x^2\,]}.$$

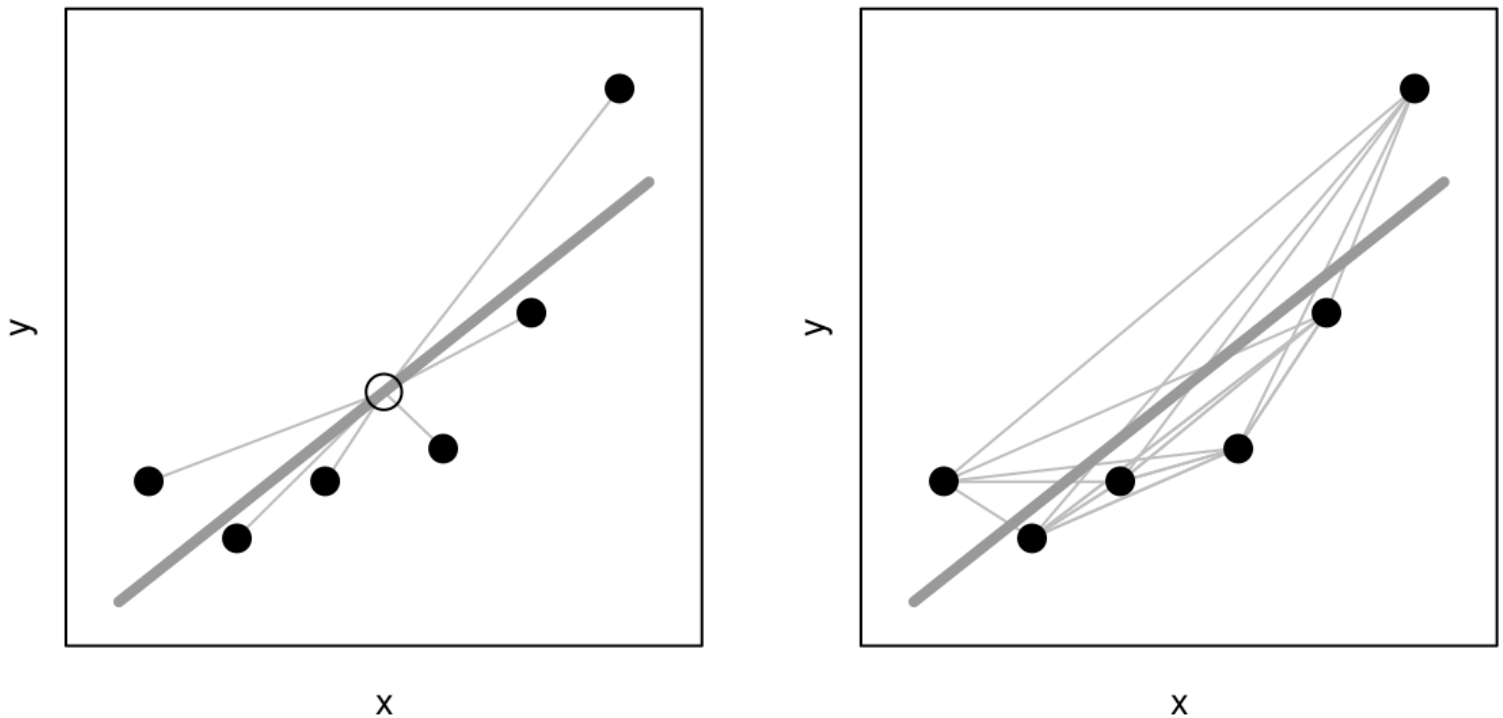
  Thus b is the case-wise slope through the origin and w its weight.

- **pairwise slopes**: For iid cases (x, y) and (x', y') we have

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

  Thus b is the pairwise slope and w its weight.

Interpretation of "slopes" in the presence of non-linearity



See Figure 5 for an illustration for samples. The formulas support the intuition that, even in the presence of nonlinearity, a linear fit can describe the overall direction of the association between the response and a regressor after adjustment.

Comments

The best approximation depends on how you measure

In the context of prediction, the objective is often to minimize a particular criterion or scoring rule. . . . [In the] case of misspecification, it is not clear which criterion should be used for estimation. In the context of forecasting conditional probabilities of binary outcomes, Elliott, Ghanem and Krüger (2016) examine this question and illustrate that the choice of scoring rule yields different best approximations to the true conditional probability function of the outcome of interest under misspecification, except under restrictive conditions. [Ghanem and Kuffner, 2019]

Predictive performance

Rinaldo et al. [2019] argue that we should give up the parameter β and instead consider:

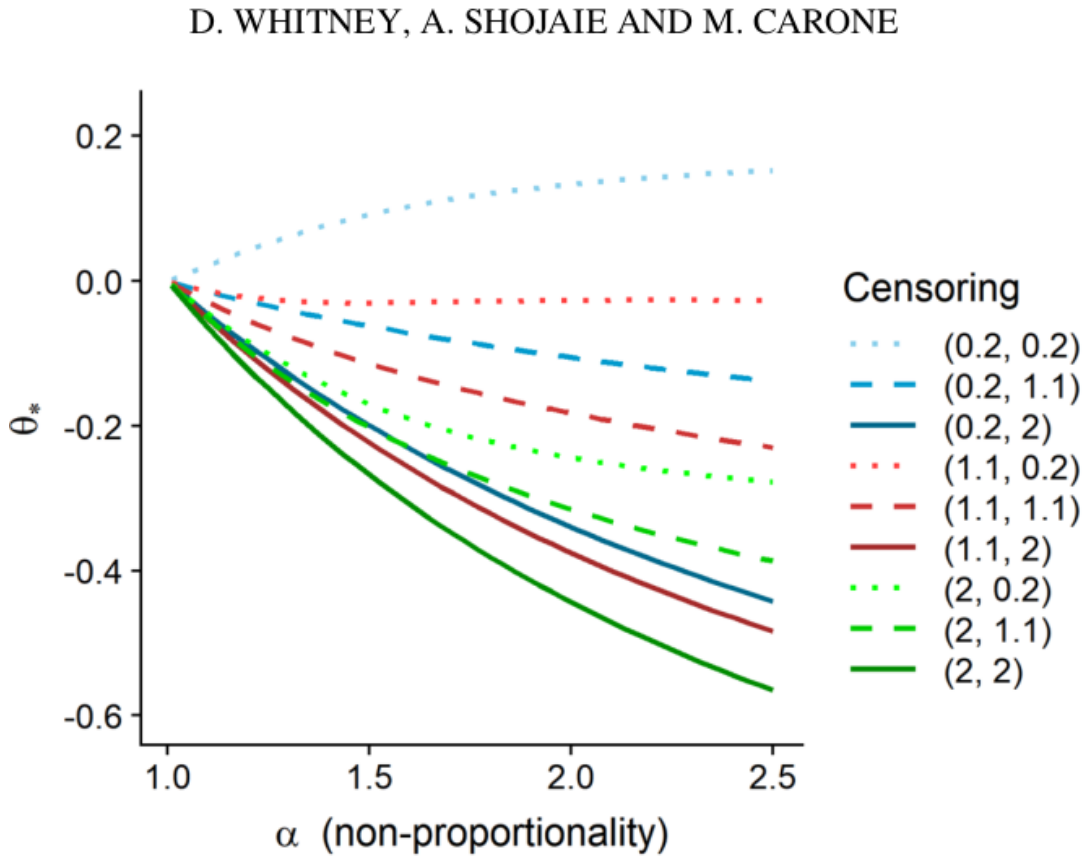
Proper causal effect

β is often mis-interpreted as a causal quantity effect. Drop the parameter β and instead define a causal quantity of interest rigorously using counterfactuals, SEMs, DAGs.

Variable importance measure

Non-parametric variable importance measures defined without reference to a model, for instance proportion of variance explained or Shapley values.

The best approximation is ill-defined when data is coarsened



The fact that the censoring distribution defines the estimand is particularly alarming. In commenting on this finding, O'Quigley (2008) states that the partial likelihood-based regression functional is not itself particularly useful nor interpretable – we agree with this viewpoint. [Whitney et al., 2019]

Fully non-parametric (model-free) parameter definition

Model $\rightarrow$ parameter (more or less interpretable)

Extend parameter from linear (or other) model to general (non-parametric) setting. The parameter interpretation simplifies to well-known quantity when the model is correct [Buja et al., 2019a,b].

Interpretable parameter $\rightarrow$ estimation by using model

Define parameter of interest directly on the non-parametric family of probability measure – model-agnostic/model-free parameter [Rinaldo et al., 2019, Whitney et al., 2019]. Separates parameter definition and estimation completely.

Flawed models as a fact of life?

Back to the main paper: In practice we are going to use some kind of estimation and approximation.

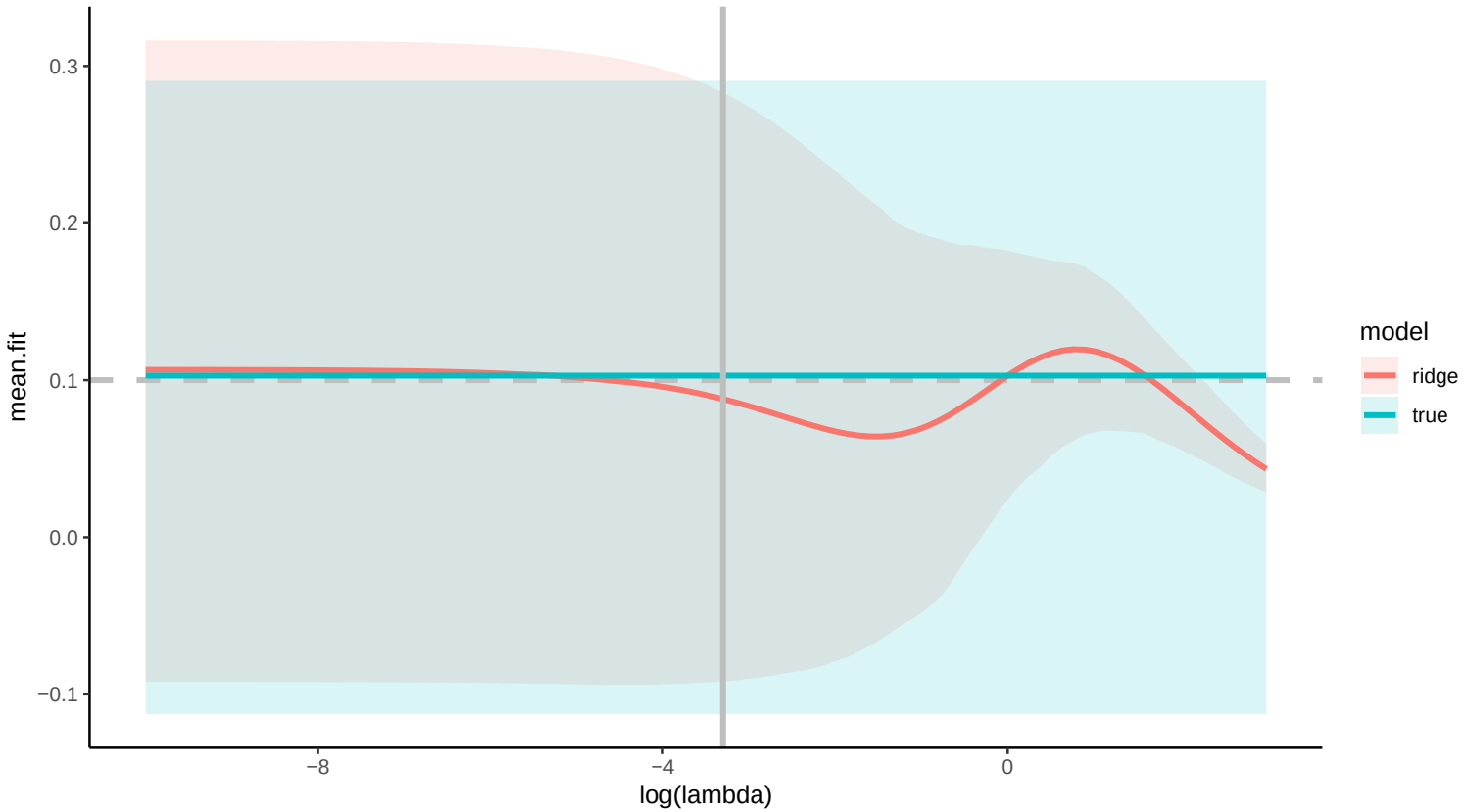
> **10. MEANINGS OF SLOPES IN THE PRESENCE OF NONLINEARITY**
>
> A first use of regressor adjustment is for proposing meanings of linear slopes in the presence of nonlinearity, and responding to Freedman's (2006, p. 302) objection: "... it is quite another thing to ignore bias [nonlinearity]. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter." Against this view one may argue that "flawed" models are a fact of life. Flaws such as nonlinearity can go undetected, or they can be tolerated for insightful simplification. A "parameter" based on best approximation is then not intrinsically wrong but in need of a useful interpretation.

The effect of estimating the nuisance parameter with an approximate nuisance model on the estimator of a *target parameter*:

- Assume the parameter of interest Ψ is identified through the nuisance parameter ν, i.e., $\Psi(P) = \tilde{\Psi}(\nu(P))$.
- If $\hat{\nu}$ is an estimator of ν, then what effect does mis-specification/approximation for the nuisance component have on the plug-in estimator $\hat{\Psi} = \tilde{\Psi}(\hat{\nu})$?

Illustration of the effect of approximate nuisance model on target estimator

Estimation of the ATE using the G-formula. For the correctly specified outcome model (blue) and a collection of mis-specified models indexed by a penalty parameter (red).



References

A. Buja, L. Brown, R. Berk, E. George, E. Pitkin, M. Traskin, K. Zhang, and L. Zhao. Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019a.

A. Buja, L. Brown, A. K. Kuchibhotla, R. Berk, E. George, and L. Zhao. Models as approximations ii: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565, 2019b.

D. Ghanem and T. A. Kuffner. Discussion: Models as approximations. *Statistical Science*, 34(4):604–605, 2019.

A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Comment: Statistical inference from a predictive perspective. *Statistical Science*, 34(4):599–603, 2019.

D. Whitney, A. Shojaie, and M. Carone. Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 34(4):591, 2019.