Why Do Statisticians Treat Predictors as Fixed? A Conspiracy Theory

Andreas Buja

joint with the PoSI Group:

Richard Berk, Lawrence Brown, Linda Zhao, Kai Zhang

Ed George, Mikhail Traskin, Emil Pitkin, Dan McCarthy

Mostly at the Department of Statistics, The Wharton School
University of Pennsylvania

WHOA-PSI — 2016/10/02

Outline

A list of mind puzzlers:

- Why are we treating regressors $\mathbf{X}$ as fixed?
- Why is the justification for conditioning on $\mathbf{X}$ wrong?
- Can we treat regressors as random?
- Degrees of misspecification are universal.
- Errors are not the only source of sampling variation.
- Model-trusting standard errors can be wrong.
- What is the meaning of regression parameters when the model is wrong?
- There exists a notion of well-specification without models.

Principle: Models, yes, but do Not Believe in Them!

Models are useful ...

- to think about data,
- to imagine generative mechanisms,
- to formalize quantities of interest (in the form of parameters),
- to guide us toward inference (testing, CIs), and
- to specify ideal conditions for inference.

Principle: Models, yes, but do Not Believe in Them!

Models are useful ...

- to think about data,
- to imagine generative mechanisms,
- to formalize quantities of interest (in the form of parameters),
- to guide us toward inference (testing, CIs), and
- to specify ideal conditions for inference.

Traditions of not believing in model correctness:

- G.E.P. Box: "......"
- Cox (1995): "The very word model implies simplification and idealization."
- A modern reason: Models are obtained by model selection.

  (Model selection makes good compromises, but does not find "truth".)

To be shown: Lip service to these maxims is not enough;
there are serious consequences.

PoSI, just briefly

- PoSI: Reduction of post-selection inference to simultaneous inference
- Computational cost: exponential in p, linear in N
  - PoSI covering all submodels feasible for $p \approx 20$.
  - Sparse PoSI for $|M| \leq 5$, e.g., feasible for $p \approx 60$.
- Coverage guarantees are marginal, not conditional.
- Statistical practice is minimally changed:
  - Use a single $\hat{\sigma}$ across all submodels M.
  - For PoSI-valid CIs, enlarge the multiplier of $\mathrm{SE}[\hat{\beta}_{j\bullet M}]$ suitably.
- PoSI covers for formal, informal, and post-hoc selection.
- PoSI solves the circularity problem of selective inference:
  Selecting regressors with PoSI tests does not invalidate the tests.
- PoSI is possible for response and transformation selection.
- Current PoSI uses fixed-**X** theory, allows 1st order misspecification,
  but assumes normality, homoskedasticity and a valid $\hat{\sigma}$.

From Fixed-**X** to Random-**X** Regression

A different but “obvious” solution: Data Splitting

Split the data into

- a model selection sample and
- an estimation & inference sample.

From Fixed-**X** to Random-**X** Regression

A different but “obvious” solution: Data Splitting

Split the data into

- a model selection sample and
- an estimation & inference sample.

Pros:

- Valid inference for the selected model.
- Works for any model
- Often less conservative than PoSI.

From Fixed-**X** to Random-**X** Regression

A different but "obvious" solution: Data Splitting

Split the data into

- a model selection sample and
- an estimation & inference sample.

Pros:

- Valid inference for the selected model.
- Works for any model
- Often less conservative than PoSI.

Cons:

- Randomness from single split
- Reduced effective sample size
- More model selection uncertainty
- More estimation uncertainty
- Loss of conditionality on **X**

Why Conditioning on **X** is Wrong

- With split-sampling and CV we break the fixed-**X** paradigm.

Why Conditioning on **X** is Wrong

- With split-sampling and CV we break the fixed-**X** paradigm.
- What is the justification for conditioning on **X**, i.e., treating **X** as fixed?

Why Conditioning on **X** is Wrong

- With split-sampling and CV we break the fixed-**X** paradigm.
- What is the justification for conditioning on **X**, i.e., treating **X** as fixed?
  Answer: Fisher's ancillarity argument for **X**.

$$\frac{p(y,x;\beta^{(1)})}{p(y,x;\beta^{(0)})} = \frac{p(y|x;\beta^{(1)})p(x)}{p(y|x;\beta^{(0)})p(x)} = \frac{p(y|x;\beta^{(1)})}{p(y|x;\beta^{(0)})}$$

Why Conditioning on **X** is Wrong

- With split-sampling and CV we break the fixed-**X** paradigm.
- What is the justification for conditioning on **X**, i.e., treating **X** as fixed?
  Answer: Fisher's ancillarity argument for **X**.

$$\frac{p(y,x;\beta^{(1)})}{p(y,x;\beta^{(0)})} = \frac{p(y|x;\beta^{(1)})p(x)}{p(y|x;\beta^{(0)})p(x)} = \frac{p(y|x;\beta^{(1)})}{p(y|x;\beta^{(0)})}$$

- Important! The ancillarity argument assumes correctness of the model.
- Refutation of Regressor Ancillarity under Misspecification:
  - Assume x_i random, $y_i = f(x_i)$ nonlinear deterministic, no errors.
    (There could be error, but this is not relevant.)
  - Fit a linear function $\hat{y}_i = \beta_0 + \beta_1 x_i$.

    $\Rightarrow$ A situation with misspecification, but no errors.

    $\Rightarrow$ Conditional on x_i, estimates $\hat{\beta}_j$ have no sampling variability.

Why Conditioning on **X** is Wrong

- With split-sampling and CV we break the fixed-**X** paradigm.
- What is the justification for conditioning on **X**, i.e., treating **X** as fixed?
  Answer: Fisher's ancillarity argument for **X**.

$$\frac{p(y, x; \beta^{(1)})}{p(y, x; \beta^{(0)})} = \frac{p(y|x; \beta^{(1)})p(x)}{p(y|x; \beta^{(0)})p(x)} = \frac{p(y|x; \beta^{(1)})}{p(y|x; \beta^{(0)})}$$

- Important! The ancillarity argument assumes correctness of the model.
- Refutation of Regressor Ancillarity under Misspecification:
  - Assume x_i random, $y_i = f(x_i)$ nonlinear deterministic, no errors.
    (There could be error, but this is not relevant.)
  - Fit a linear function $\hat{y}_i = \beta_0 + \beta_1 x_i$.

    $\Rightarrow$ A situation with misspecification, but no errors.

    $\Rightarrow$ Conditional on x_i, estimates $\hat{\beta}_j$ have no sampling variability.

However, watch **"The Unconditional Movie"!**

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

Random **X** and Misspecification

Randomness of X and misspecification conspire to create sampling variability in the estimates.

- Consequence: Under misspecification, conditioning on **X** is wrong.
- Source 1 of model-robust & unconditional inference:

  Asymptotic inference based on Eicker/Huber/White's

  Sandwich Estimator of Standard Error.

$$\mathbf{V}[\hat{\beta}] \;=\; \mathbf{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}\;\mathbf{E}[(Y-\vec{\boldsymbol{X}}'\,\beta)\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]\;\;\mathbf{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'\,]^{-1}.$$

- Source 2 of model-robust & unconditional inference:

  the Pairs or x-y Bootstrap (not the residual bootstrap)

- Validity of inference is in an assumption-lean/model-robust framework...

An Assumption-Lean Framework

- $\exists$ joint distribution, i.i.d. sampling: $(y_i, \vec{\boldsymbol{x}}_i) \sim \mathrm{P}(dy, d\vec{\boldsymbol{x}}) = \mathrm{P}_{Y,\vec{\boldsymbol{X}}}$

  Assume properties sufficient to grant CLTs for estimates of interest.

- !!! No assumptions on $\mu(\vec{\boldsymbol{x}}) = \mathbf{E}[Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}]$, $\sigma^2(\vec{\boldsymbol{x}}) = \mathbf{V}[Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}]$!!!

- Define a population OLS parameter:

$$\boldsymbol{\beta} := \mathrm{argmin}_{\tilde{\boldsymbol{\beta}}} \, \mathbf{E}\left[\left(Y - \tilde{\boldsymbol{\beta}}' \vec{\boldsymbol{X}}\right)^2\right] = \mathbf{E}[\vec{\boldsymbol{X}} \vec{\boldsymbol{X}}']^{-1} \, \mathbf{E}[\vec{\boldsymbol{X}} Y]$$

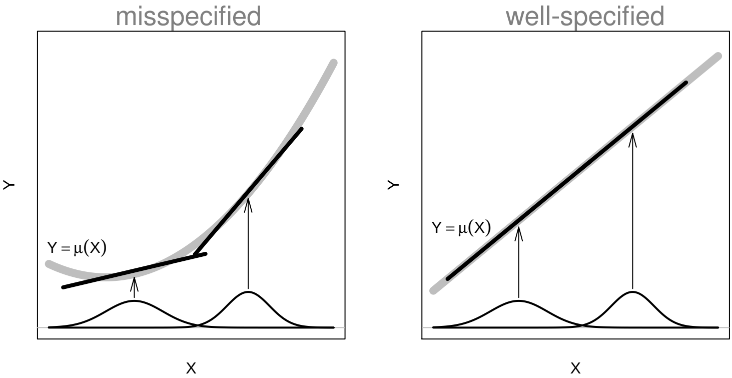
- This is the target of inference: $\boxed{\boldsymbol{\beta} = \boldsymbol{\beta}(\mathrm{P})}$

  $\Rightarrow$ $\boldsymbol{\beta}$ is a statistical functional — a **“Regression Functional”.**

“Statistical Functional” View of Regression

(“Random **X** Theory”)

Implication 1 of Misspecification/Model-Robustness



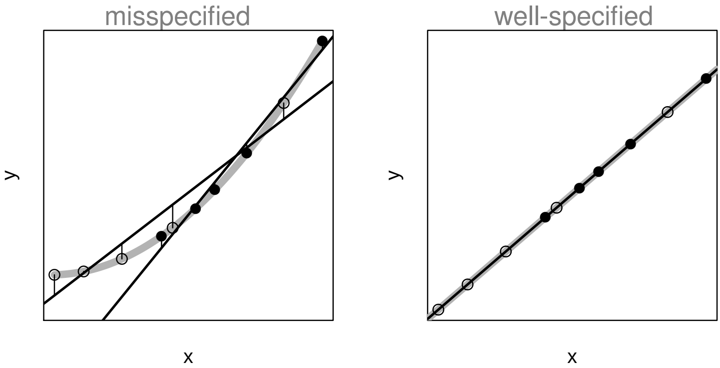
Under misspecification parameters depend on the $\vec{X}$ distribution:

$$\beta(\mathrm{P}_{Y,\vec{X}}) \;=\; \beta(\mathrm{P}_{Y|\vec{X}}, \mathrm{P}_{\vec{X}}) \qquad\qquad \beta(\mathrm{P}_{Y,\vec{X}}) \;=\; \beta(\mathrm{P}_{Y|\vec{X}})$$

Upside down: “Ancillarity = parameters do not affect the distribution.”

Upside up: "Ancillarity = the parameters are not affected by the distribution.”

Implication 2 of Misspecification/Model-Robustness



Misspecification and random $\vec{X}$ generate sampling variability:

$\mathbf{V}[\,\mathbf{E}[\,\hat{\beta}|\mathbf{X}\,]\,] > \mathbf{0}$ $\qquad$ $\mathbf{V}[\,\mathbf{E}[\,\hat{\beta}|\mathbf{X}\,]\,] = \mathbf{0}$

Recall "the unconditional movie."

Mis/Well-Specification of Regression Functionals

A powerful extension of the idea of mis/well-specification:

- Write a regression functional as $\boldsymbol{\beta}(\mathrm{P}) = \boldsymbol{\beta}(\mathrm{P}_{Y|\vec{\boldsymbol{X}}}, \mathrm{P}_{\vec{\boldsymbol{X}}})$.
- Definition: A regression functional $\boldsymbol{\beta}(\mathrm{P})$ is well-specified for $\mathrm{P}_{Y|\vec{\boldsymbol{X}}}$ if

$$\boldsymbol{\beta}(\mathrm{P}_{Y|\vec{\boldsymbol{X}}}, \mathrm{P}_{\vec{\boldsymbol{X}}}) \;=\; \boldsymbol{\beta}(\mathrm{P}_{Y|\vec{\boldsymbol{X}}})$$

- Under well-specification ...
  - $\boldsymbol{\beta}(\mathrm{P})$ does not depend on the $\vec{\boldsymbol{X}}$ distribution, and
  - $\mathbf{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]$ does not have sampling variability due to conspiracy.

  $\Rightarrow$ A new form of regressor ancillarity

- OLS: $\boldsymbol{\beta}(\mathrm{P})$ is well-specified for P iff $\mathbf{E}[Y|\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}] = \boldsymbol{\beta}' \vec{\boldsymbol{x}}$.

Sampling Variability of Regression Functionals

Plug-in estimates $\hat{\beta} = \beta(\hat{\mathrm{P}})$ of regression functionals (such as OLS estimators) have two sources of variability:

- the conditional distribution of $\mathbf{y}|\mathbf{X}$,
- the marginal distribution of $\mathbf{X}$ combined with misspecification.

This can be expressed with the formula

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

$\mathbf{V}[\,\hat{\beta}\,|\,\mathbf{X}\,]$: The only source of sampling variability in linear models theory.

$\mathbf{E}[\,\hat{\beta}\,|\,\mathbf{X}\,]$: The target of estimation in linear models theory,
but really a random variable under misspecification,
hence a source of sampling variability.

$\Rightarrow$ $\mathbf{V}[\;\mathbf{E}[\,\hat{\beta}\,|\,\mathbf{X}\,]$ is the "**conspiracy** part of sampling variability,
caused by a synergy of randomness of **X** and misspecification.

CLTs under Misspecification: Sandwich Form

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

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

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

$\mu(\vec{\boldsymbol{X}}) = \mathbf{E}[Y|\vec{\boldsymbol{X}}]$			response surface
$\eta(\vec{\boldsymbol{X}}) = \mu(\vec{\boldsymbol{X}}) - \beta(\mathrm{P})'\,\vec{\boldsymbol{X}}$			nonlinearity
$\sigma^2(\vec{\boldsymbol{X}}) = \mathbf{V}[Y|\vec{\boldsymbol{X}}]$			conditional noise variance
$\epsilon = Y - \mu(\vec{\boldsymbol{X}}),$			noise
$\delta = Y - \beta(\mathrm{P})'\,\vec{\boldsymbol{X}}$	$= \eta(\vec{\boldsymbol{X}}) + \epsilon,$		population residual
$m^2(\vec{\boldsymbol{X}}) = \mathbf{E}[\,\delta^2\,|\,\vec{\boldsymbol{X}}]$	$= \eta^2(\vec{\boldsymbol{X}}) + \sigma^2(\vec{\boldsymbol{X}})$		conditional MSE

The x-y Bootstrap for Regression

- Model-robust framework: The nonparametric x-y bootstrap applies.

  $$\text{Resample } (\vec{\boldsymbol{x}}_i, y_i) \text{ pairs} \quad \rightarrow \quad (\vec{\boldsymbol{x}}_i^*, y_i^*) \quad \rightarrow \quad \hat{\beta}^*.$$

  Estimate $\mathrm{SE}(\hat{\beta}_j)$ by $\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\beta}_j) = \mathrm{SD}^*(\beta_j^*)$.

- Let $\hat{\mathrm{SE}}_{\mathrm{lin}}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\|\mathbf{x}_{j\bullet}\|}$ be the usual standard error erstimate of linear models theory .

- **Question:** Is the following always true?

  $$\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\beta}_j) \overset{?}{\approx} \hat{\mathrm{SE}}_{\mathrm{lin}}(\hat{\beta}_j)$$

  Compare conventional and bootstrap standard errors empirically...

Conventional vs Bootstrap Std Errors

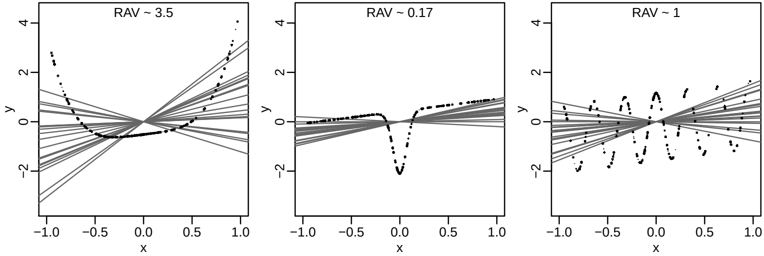
- LA Homeless Data (Richard Berk, UPenn)
- Response: `StreetTotal` of homeless in a census tract, $N = 505$
- $R^2 \approx 0.13$, residual dfs $= 498$

	$\hat{\beta}_j$	$\mathrm{SE}_{\mathrm{lin}}$	$\mathrm{SE}_{\mathrm{boot}}$	$\mathrm{SE}_{\mathrm{boot}}/\mathrm{SE}_{\mathrm{lin}}$	t_{lin}
MedianInc	-4.241	4.342	2.651	**0.611**	-0.977
PropVacant	18.476	3.595	5.553	**1.545**	5.140
PropMinority	2.759	3.935	3.750	0.953	0.701
PerResidential	-1.249	4.275	2.776	**0.649**	-0.292
PerCommercial	10.603	3.927	5.702	**1.452**	2.700
PerIndustrial	11.663	4.139	7.550	**1.824**	2.818

- Reason for the discrepancy: misspecification.

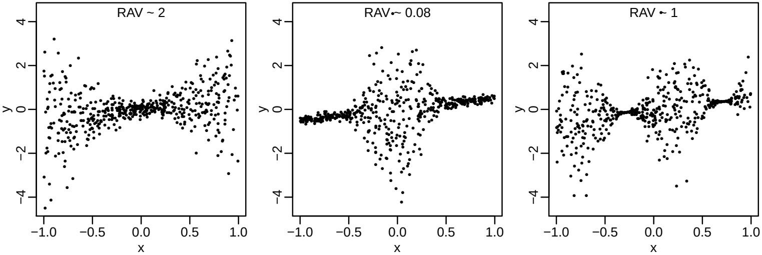
Reason 1 for $\mathrm{SE}_{\mathrm{boot}} \neq \mathrm{SE}_{\mathrm{lin}}$: Nonlinearity

Which has the smallest/largest true $\mathrm{SE}(\hat{\beta})$?



Reason 2 for $\mathrm{SE}_{\mathrm{boot}} \neq \mathrm{SE}_{\mathrm{lin}}$: Heteroskedasticity

Which has the smallest/largest true $\mathrm{SE}(\hat{\beta})$?



Sandwich and *x*-*y* Bootstrap Estimators

- The *x*-*y* bootstrap is asymptotically correct in the assumption-lean/model-robust framework,
  and so is the sandwich estimator of standard error.

- There exists a connection, based on the *M*-of-*N* bootstrap:
  Resample datasets of size *M* out of *N* with replacement and rescale

  $$\hat{\mathrm{SE}}_{M:N}[\hat{\beta}_j] \;:=\; (M/N)^{1/2}\,\mathrm{SD}^*[\hat{\beta}_j^*]$$

  Proposition: $\hat{\mathrm{SE}}_{M:N}[\hat{\beta}_j] \;\to\; \hat{\mathrm{SE}}_{sand}[\hat{\beta}_j] \quad (M \to \infty)$

  I.e., the sandwich estimator is the limit of the *M*-of-*N* boostrap estimator.

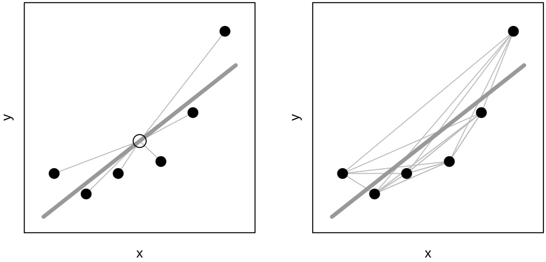
- This holds for all sandwich estimators, not just those of OLS.

- Bootstrap estimators for small *M* may have advantages.

The Meaning of Slopes under Misspecification

Allowing misspecification messes up our regression practice:

What is the meaning of slopes under misspecification?



Case-wise slopes: $\hat{\beta} = \sum_i w_i \, b_i \,, \quad b_i := \frac{y_i}{x_i}, \quad w_i := \frac{x_i^2}{\sum_{i'} x_{i'}^2}$

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

Conclusions

- Use models, but don't believe in them.
- Main use of models: Defining parameters.
- Extend the parameters beyond the model $\rightarrow$ regression functionals.
- There is a new notion of well-specification for regression functionals.
- Random $\mathbf{X}$ & model misspecification generate sampling variability.
- Use inference that does not assume model correctness.
- Outlook: PoSI under complete misspecification

Conclusions

- Use models, but don't believe in them.
- Main use of models: Defining parameters.
- Extend the parameters beyond the model $\rightarrow$ regression functionals.
- There is a new notion of well-specification for regression functionals.
- Random $\mathbf{X}$ & model misspecification generate sampling variability.
- Use inference that does not assume model correctness.
- Outlook: PoSI under complete misspecification

THANKS!