Post-Selection Inference for Models that are Approximations

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

Rutgers University — 2016/04/20

Problems: Non-Reproducibility in Biomedical Science

Borrowed from: Berger, 2012, "Reproducibility of Science: P-values and Multiplicity"

- Bayer Healthcare: 67 attempts at replicating published research findings
  - Fewer than $1/4$ were viewed as replicated.
  - Over 2/3 had major inconsistencies leading to project termination.
- Arrowsmith (2011, Nat.Rev.Drug Discovery 10): Drug trial success rates $\downarrow$
  - Phase II: 28% in 2005, 18% in 2010
  - Phase III: 80% in 2000, 50% in 2010
  - Phase III cancer drugs: 30%
- The NIH funded randomized clinical trials to follow up exciting results from 20 observational studies: Only one was replicated.
- Ioannidis (JAMA-2005, 218-28):
  - 5 of 6 highly cited nonrandomized studies were contradicted or had found stronger effects than were established by later studies.

Most Empirical Findings Are False

Bombshell in Biomedical Science:

- **"Why Most Published Research Findings Are False"**
  by Ioannidis (2005, PLOS Medicine)
- Demonstrates the combined influences of: Pre-study (true/false) odds R, Type I error α, power β, bias u (!), # independent similar studies n.
- Famous corollaries:
  - the smaller the study sizes,
  - the smaller the effect sizes,
  - the greater the number of tested relationships,
  - the greater the flexibility in design, definitions, outcomes, techniques,
  - the greater the financial and professional interests,
  - the hotter the field,

  the less likely the research findings are to be true.
- Note: **"bias"** due to "manipulation in the analysis", "selective reporting"

Problems: “False-Positive” Social Sciences

Bombshell in psychological research:

- **“False Positive Psychology: Undisclosed Flexiblity in Data Collection and Analysis allows Presenting Anything as Significant”** by Simmons, Nelson, Simonsohn (2011, Psychological Science)
- New concept: “**Researcher Degrees of Freedom**”
  - “In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?”
  - Elaboration of what Ioannidis could have meant with “bias”.
- Soul-searching in social science journals:
  - disclosure requirements, emphasis on replication, ...

Problem: “False-Positive” Social Sciences (contd.)

- From Simmons, Nelson, Simonsohn (2011):

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: stepwise, all-subsets with AIC, BIC,...; Lasso; Dantzig;...

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: stepwise, all-subsets with AIC, BIC,...; Lasso; Dantzig;...
  - informal selection: residual plots, influence diagnostics, ...

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: stepwise, all-subsets with AIC, BIC,...; Lasso; Dantzig;...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: “This predictor is too costly given its effect size.”

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: stepwise, all-subsets with AIC, BIC,...; Lasso; Dantzig;...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: "This predictor is too costly given its effect size."

- Suspicions:
  - All three modes of model selection may be used in much empirical research.
  - Ironically, the most thorough and competent data analysts may also be the ones who produce the most spurious findings.
  - Post-selection inference for "adaptive Lasso", say, won't solve the problem: Few empirical researchers commit themselves

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: stepwise, all-subsets with AIC, BIC,...; Lasso; Dantzig;...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: “This predictor is too costly given its effect size.”

- Suspicions:
  - All three modes of model selection may be used in much empirical research.
  - Ironically, the most thorough and competent data analysts may also be the ones who produce the most spurious findings.
  - Post-selection inference for “adaptive Lasso”, say, won’t solve the problem: Few empirical researchers commit themselves

    a priori

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: stepwise, all-subsets with AIC, BIC,...; Lasso; Dantzig;...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: "This predictor is too costly given its effect size."

- Suspicions:
  - All three modes of model selection may be used in much empirical research.
  - Ironically, the most thorough and competent data analysts may also be the ones who produce the most spurious findings.
  - Post-selection inference for "adaptive Lasso", say, won't solve the problem: Few empirical researchers commit themselves

    a priori to one formal selection method

Statistical Biases – One Among Several

- Consider now one cause of statistical bias:

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: stepwise, all-subsets with AIC, BIC,...; Lasso; Dantzig;...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: "This predictor is too costly given its effect size."

- Suspicions:
  - All three modes of model selection may be used in much empirical research.
  - Ironically, the most thorough and competent data analysts may also be the ones who produce the most spurious findings.
  - Post-selection inference for "adaptive Lasso", say, won't solve the problem: Few empirical researchers commit themselves

    a priori to one formal selection method and nothing else.

Linear Model Inference and Variable Selection

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ = fixed design matrix, $N \times p$, $N > p$, full rank.
- $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$

In textbooks:

1. Variables selected
2. Data seen
3. Inference produced

In common practice:

1. Data seen
2. Variables selected
3. Inference produced

Linear Model Inference and Variable Selection

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ = fixed design matrix, $N \times p$, $N > p$, full rank.
- $\epsilon \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$

In textbooks:

1. Variables selected
2. Data seen
3. Inference produced

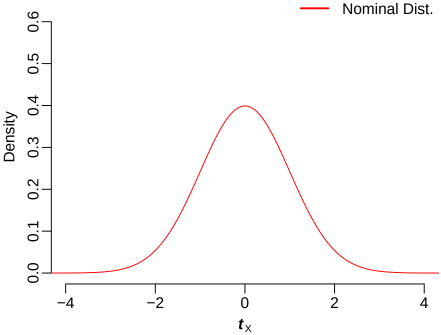
In common practice:

1. Data seen
2. Variables selected
3. Inference produced

Is this inference valid?

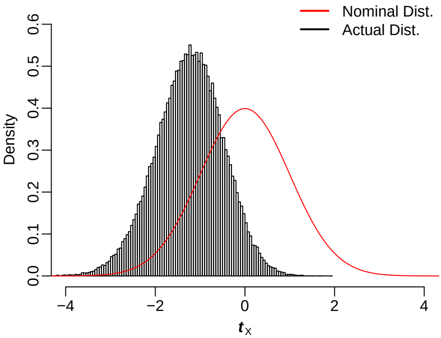
Evidence from a Simulation

Marginal Distribution of Post-Selection *t*-statistics:



Evidence from a Simulation

Marginal Distribution of Post-Selection *t*-statistics:



- The overall coverage probability of the conventional post-selection CI is $83.5\% < 95\%$.
- For $p = 30$, the coverage probability can be as low as 39%.

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.
- The approach is a reduction of PoSI to simultaneous inference.
- Simultaneity is across all submodels and all slopes in them.
- As a result, we obtain

valid PoSI for all variable selection procedures!

The PoSI Procedure — Rough Outline

- We propose to construct Post Selection Inference (PoSI) with guarantees for the coverage of CIs and Type I errors of tests.
- We widen CIs and retention intervals to achieve correct/conservative post-selection coverage probabilities. This is the price we have to pay.
- The approach is a reduction of PoSI to simultaneous inference.
- Simultaneity is across all submodels and all slopes in them.
- As a result, we obtain

  valid PoSI for all variable selection procedures!

- But first we need make sense of

  Targets of Inference in Approximate Models

Incorrect Submodels — What Is Being Estimated?

- Denote a submodel by integers $\mathrm{M} = \{j_1, j_2, ..., j_m\}$:

$$\mathbf{X}_{\mathrm{M}} = (\mathbf{X}_{j_1}, \mathbf{X}_{j_2}, ..., \mathbf{X}_{j_m}) \in \mathbb{R}^{N \times m}.$$

- OLS coefficient estimates in the submodel M:

$$\hat{\beta}_{\mathrm{M}} = (\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \mathbf{Y} \in \mathbb{R}^m$$

- Q: What does $\hat{\beta}_{\mathrm{M}}$ estimate, **not** assuming the truth of M?

Incorrect Submodels — What Is Being Estimated?

- Denote a submodel by integers $\mathrm{M} = \{j_1, j_2, ..., j_m\}$:

$$\mathbf{X}_{\mathrm{M}} = (\mathbf{X}_{j_1}, \mathbf{X}_{j_2}, ..., \mathbf{X}_{j_m}) \in \mathbb{R}^{N \times m}.$$

- OLS coefficient estimates in the submodel M:

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = (\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \mathbf{Y} \in \mathbb{R}^m$$

- Q: What does $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ estimate, **not** assuming the truth of M?
  A: Its expectation!

$$\boldsymbol{\mu} := \mathbf{E}[\mathbf{Y}] \in \mathbb{R}^N \quad \textbf{arbitrary!!}$$

$$\boldsymbol{\beta}_{\mathrm{M}} := \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}] = (\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \boldsymbol{\mu}$$

We do **not** assume that the submodel is correct: $\boldsymbol{\mu} \neq \mathbf{X}_{\mathrm{M}} \boldsymbol{\beta}_{\mathrm{M}}$ allowed!
But $\mathbf{X}_{\mathrm{M}} \boldsymbol{\beta}_{\mathrm{M}}$ is the best approximation to $\boldsymbol{\mu}$.

Adjustment, Estimates, Parameters, t-Statistics

Notation and facts for the components of $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\boldsymbol{\beta}_{\mathrm{M}}$, assuming $j \in \mathrm{M}$:

- Let $\mathbf{X}_{j\bullet\mathrm{M}}$ be the predictor $\mathbf{X}_j$ adjusted for the other predictors in M:

$$\mathbf{X}_{j\bullet\mathrm{M}} \;:=\; \left(\mathbf{I} - \mathbf{H}_{\mathrm{M}\setminus\{j\}}\right) \mathbf{X}_j \;\;\perp\;\; \mathbf{X}_k \;\forall k \in \mathrm{M} \setminus \{j\}.$$

- Let $\hat{\beta}_{j\bullet\mathrm{M}}$ be the slope estimate and $\beta_{j\bullet\mathrm{M}}$ be the parameter for $\mathbf{X}_j$ in M:

$$\hat{\beta}_{j\bullet\mathrm{M}} \;:=\; \frac{\langle \mathbf{X}_{j\bullet\mathrm{M}}, \mathbf{Y} \rangle}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|^2}, \qquad \beta_{j\bullet\mathrm{M}} \;:=\; \frac{\langle \mathbf{X}_{j\bullet\mathrm{M}}, \mathbf{E}[\mathbf{Y}] \rangle}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|^2}.$$

- Let $t_{j\bullet\mathrm{M}}$ be the t-statistic for $\hat{\beta}_{j\bullet\mathrm{M}}$ and $\beta_{j\bullet\mathrm{M}}$:

$$t_{j\bullet\mathrm{M}} \;:=\; \frac{\hat{\beta}_{j\bullet\mathrm{M}} - \beta_{j\bullet\mathrm{M}}}{\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|} \;=\; \frac{1}{\hat{\sigma}} \left\langle \frac{\mathbf{X}_{j\bullet\mathrm{M}}^T}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|}, \mathbf{Y} - \mathbf{E}[\mathbf{Y}] \right\rangle.$$

Parameters One More Time

- Important: If the predictors are partly collinear (non-orthogonal) then

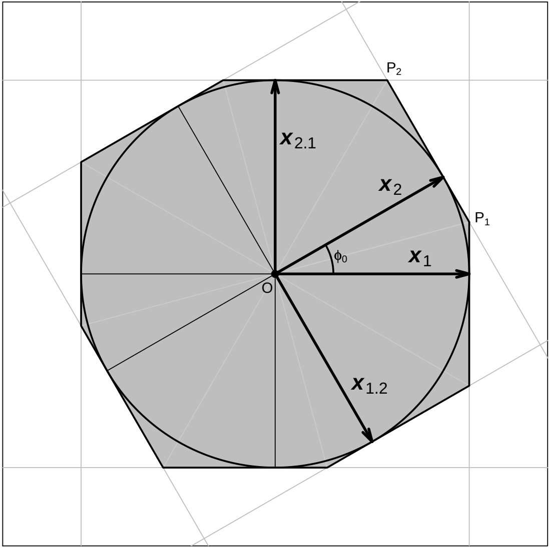
$$\boxed{\mathrm{M} \neq \mathrm{M}', \;\; j \in M \cap M' \quad \Rightarrow \quad \beta_{j\bullet\mathrm{M}} \neq \beta_{j\bullet\mathrm{M}'}}$$

in value and in meaning!

- Rule: A difference in adjustment implies a difference in parameters.

- Number of parameters $\beta_{j\bullet\mathrm{M}}$: $p\,2^{p-1}$

Geometry of Adjustment



Column space of **X** for $p = 2$ predictors, partly collinear

Error Estimates $\hat{\sigma}$: One for All Submodels

- Critical Point: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,
  use **one error estimate** $\hat{\sigma}$ for all submodels.
  - Do **not** use ~~$\hat{\sigma}_{\mathrm{M}}$~~ !
  - Use $\hat{\sigma} = \hat{\sigma}_{Full}$ instead for all submodels M.
  - $t_{j\bullet\mathrm{M}}$ will have a t-distribution with the same dfs $\forall\mathrm{M}, \ \forall j \in \mathrm{M}$.

- Q: What if even the full model is 1st order wrong?

Error Estimates $\hat{\sigma}$: One for All Submodels

- Critical Point: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,
  use **one error estimate** $\hat{\sigma}$ for all submodels.
  - Do **not** use ~~$\hat{\sigma}_{\mathrm{M}}$~~ !
  - Use $\hat{\sigma} = \hat{\sigma}_{Full}$ instead for all submodels M.
  - $t_{j\bullet\mathrm{M}}$ will have a t-distribution with the same dfs $\forall\mathrm{M},\ \forall j \in \mathrm{M}$.

- Q: What if even the full model is 1st order wrong?
  A: $\hat{\sigma}_{Full}$ will be inflated and inference will be conservative.

- A better $\hat{\sigma}$ is available if ...
  - exact replicates exist: use $\hat{\sigma}$ from the 1-way ANOVA of replicates;
  - a larger than the full model can be assumed 1st order correct: use $\hat{\sigma}_{Large}$;
  - a previous dataset provided a valid estimate: use $\hat{\sigma}_{previous}$;
  - nonparametric estimates are available: use $\hat{\sigma}_{nonpar}$ (Hall and Carroll 1989).

Statistical Inference under First Order Incorrectness

- Same correct inference across all submodels M and all $\beta_{j \bullet \mathrm{M}}$:
  - If $r =$ dfs in $\hat{\sigma}$ and $K = t_{1-\alpha/2,r}$, then the "almost usual" interval

$$\mathrm{CI}_{j \bullet \mathrm{M}}(K) := [\, \hat{\beta}_{j \bullet \mathrm{M}} \pm K \hat{\sigma} / \| \mathbf{X}_{j \bullet \mathrm{M}} \| \,]$$

satisfies:

$$\mathbf{P}[\, \beta_{j \bullet \mathrm{M}} \in \mathrm{CI}_{j \bullet \mathrm{M}}(K) \,] = 1 - \alpha \quad \forall \mathrm{M}, \ \forall j \in \mathrm{M}$$

Statistical Inference under First Order Incorrectness

- Same correct inference across all submodels M and all $\beta_{j\bullet\mathrm{M}}$:
  - If $r =$ dfs in $\hat{\sigma}$ and $K = t_{1-\alpha/2,r}$, then the "almost usual" interval

$$\mathrm{CI}_{j\bullet\mathrm{M}}(K) \;:=\; [\,\hat{\beta}_{j\bullet\mathrm{M}} \pm K\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|\,]$$

satisfies:

$$\mathbf{P}[\,\beta_{j\bullet\mathrm{M}} \in \mathrm{CI}_{j\bullet\mathrm{M}}(K)\,] \;=\; 1-\alpha \quad \forall\mathrm{M},\ \forall j \in \mathrm{M}$$

- Correct inference in a mean-misspecified homoskedastic model:

$$\mathbf{Y} \;=\; \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I})$$

  - Permitted: $\boldsymbol{\mu} \neq \mathbf{X}\boldsymbol{\beta},\ \mathbf{X}_{\mathrm{M}}\boldsymbol{\beta}_{\mathrm{M}}\ \forall\mathrm{M}$
  - A single valid $\hat{\sigma}$ with known dfs across all submodels enables **simultaneous inference** across submodels.

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on $\mathbf{X}$ does not invalidate inference (example: deselect predictors based on VIF, $\mathbf{H}$, ...).

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y}), \quad \mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on $\mathbf{X}$ does not invalidate inference (example: deselect predictors based on VIF, $\mathbf{H}$, ...).

- Candidates for meaningful coverage probabilities:

$$\mathbf{P}[\ \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\ ]$$

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on $\mathbf{X}$ does not invalidate inference (example: deselect predictors based on VIF, $\mathbf{H}$, ...).

- Candidates for meaningful coverage probabilities:

  $\mathbf{P}[\ \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\ ]$ Meaningless!

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\quad \mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Candidates for meaningful coverage probabilities:

$$\mathbf{P}[\ \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K) \mid \hat{\mathrm{M}} = \mathrm{M}\ ] \quad \forall j \in \mathrm{M} \quad \text{(Taylor et al.)}$$

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Candidates for meaningful coverage probabilities:

$$\mathbf{P}[\ \forall j \in \hat{\mathrm{M}} : \ \beta_{j \bullet \hat{\mathrm{M}}} \in \mathrm{CI}_{j \bullet \hat{\mathrm{M}}}(K)\ ]$$

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \rightarrow \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Candidates for meaningful coverage probabilities:

$$\mathbf{P}[\ \forall j \in \hat{\mathrm{M}} :\ \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\ ]$$

- Larger Problem: Which analyst will commit to one $\hat{\mathrm{M}}$ and no other?

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\quad \mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Candidates for meaningful coverage probabilities:

$$\mathbf{P}[\ \forall j \in \hat{\mathrm{M}} : \ \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\ ]$$

- Larger Problem: Which analyst will commit to one $\hat{\mathrm{M}}$ and no other?

- Solution: Ask for more!

Variable Selection

- What is a variable selection procedure?

  A map $\mathbf{Y} \mapsto \hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, $\mathbb{R}^N \to \mathcal{P}(\{1, ...p\})$

  - $\hat{\mathrm{M}}$ divides the response space $\mathbb{R}^N$ into up to 2^p subsets.
  - In a fixed-predictor framework, selection purely based on **X** does not invalidate inference (example: deselect predictors based on VIF, **H**, ...).

- Candidates for meaningful coverage probabilities:

$$\mathbf{P}[\ \forall j \in \hat{\mathrm{M}} :\ \beta_{j\bullet\hat{\mathrm{M}}} \in \mathrm{CI}_{j\bullet\hat{\mathrm{M}}}(K)\ ]$$

- Larger Problem: Which analyst will commit to one $\hat{\mathrm{M}}$ and no other?

- Solution: Ask for more!

  Universal Post-Selection Inference for all selection procedures

Reduction to Simultaneous Inference

Lemma

For any variable selection procedure $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, *we have the following "significant triviality bound":*

$$\max_{j \in \hat{\mathrm{M}}} |t_{j \bullet \hat{\mathrm{M}}}| \;\leq\; \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j \bullet \mathrm{M}}| \qquad \forall\, \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^N.$$

Reduction to Simultaneous Inference

Lemma

For any variable selection procedure $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, *we have the following "significant triviality bound":*

$$\max_{j \in \hat{\mathrm{M}}} |t_{j\bullet\hat{\mathrm{M}}}| \;\leq\; \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j\bullet\mathrm{M}}| \qquad \forall\, \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^N.$$

Theorem

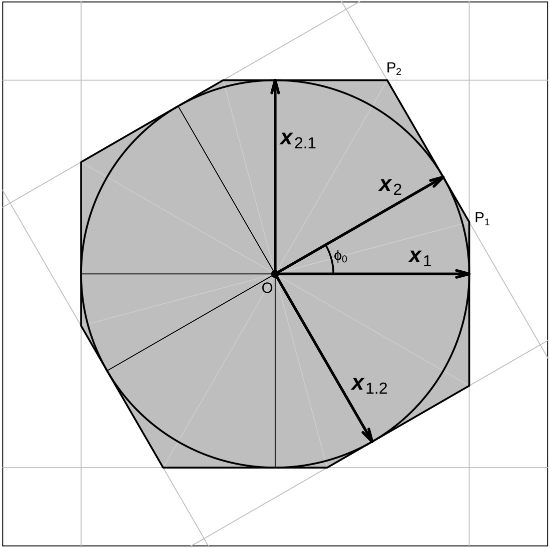
Let K *be the* $1-\alpha$ *quantile of the* "**max-max-**$|t|$" *statistic of the lemma:*

$$\mathbf{P}\big[\, \max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j\bullet\mathrm{M}}| \leq K \,\big] \overset{(\geq)}{=} 1 - \alpha.$$

Then we have the following universal PoSI guarantee:

$$\mathbf{P}\big[\, \beta_{j\bullet\hat{\mathrm{M}}} \in CI_{j\bullet\hat{\mathrm{M}}}(K) \;\; \forall j \in \hat{\mathrm{M}} \,\big] \;\geq\; 1 - \alpha \qquad \forall \hat{\mathrm{M}}.$$

PoSI Geometry — Simultaneity



PoSI polytope
= intersection
of all *t*-bands.

Computing PoSI

- The simultaneity challenge: $\#\{|t_{j\bullet\mathrm{M}}|\} = p\,2^{p-1}$

p	3	4	5	6	7	8	9	10	11
$\#\|t\|$	12	32	80	192	448	1, 024	2, 304	5, 120	11, 264
p	12	13	14	15	16	17	18	19	20
$\#\|t\|$	24, 576	53, 248	114, 688	245, 760	524, 288	1, 114, 112	2, 359, 296	4, 980, 736	10, 485, 760

Computing PoSI

- The simultaneity challenge: $\#\{|t_{j\bullet\mathrm{M}}|\} = p\,2^{p-1}$

p	3	4	5	6	7	8	9	10	11
$\#\|t\|$	12	32	80	192	448	1,024	2,304	5,120	11,264
p	12	13	14	15	16	17	18	19	20
$\#\|t\|$	24,576	53,248	114,688	245,760	524,288	1,114,112	2,359,296	4,980,736	10,485,760

- **Computations**: (for R Code, search "Buja Wharton")
  - Computational cost is linear in N, exponential in p.
  - Off-the-shelf R software works up to $p \approx 7$.
  - Custom semi-MC-approximation in R works up to $p \approx 20$.
  - Sparse PoSI: Limit search to models of size $\leq m$; permit $N < p$, $m \leq N$.
    Example: PoSI for $p = 50$ and $m = 5$ requires $\#\{|t_{j\bullet\mathrm{M}}|\}$ = 11,576,300.

Computing PoSI

- The simultaneity challenge: $\#\{|t_{j\bullet\mathrm{M}}|\} = p\,2^{p-1}$

p	3	4	5	6	7	8	9	10	11
$\#\|t\|$	12	32	80	192	448	1, 024	2, 304	5, 120	11, 264
p	12	13	14	15	16	17	18	19	20
$\#\|t\|$	24, 576	53, 248	114, 688	245, 760	524, 288	1, 114, 112	2, 359, 296	4, 980, 736	10, 485, 760

- **Computations**: (for R Code, search "Buja Wharton")
  - Computational cost is linear in N, exponential in p.
  - Off-the-shelf R software works up to $p \approx 7$.
  - Custom semi-MC-approximation in R works up to $p \approx 20$.
  - Sparse PoSI: Limit search to models of size $\leq m$; permit $N < p$, $m \leq N$.
    Example: PoSI for $p = 50$ and $m = 5$ requires $\#\{|t_{j\bullet\mathrm{M}}|\}$ = 11,576,300.
- **Large-*p* Asymptotics**: based on sequences of structured designs **X**
  - Worst-case: $K(p) \in \sqrt{p} \cdot [0.78, 0.866...]$
  - Best-case : $K(p) \sim \sqrt{2\log(p)}$

PoSI Benefits

PoSI protection may seem conservative, **but**

PoSI inference will be valid even if one...

... tries several formal selection methods and picks the “best”;

... uses informal model diagnostics to reject models;

... performs “significance hunting”, i.e., selects the model with the most significant effects on preferred predictors;

... steps forward/backward till all selected predictors are “significant”;

... analyzes clinical trial data in post-hoc “data mining”.

PoSI from Split Samples

Very different "obvious" approach: Split the data into

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

PoSI from Split Samples

Very different “obvious” approach: Split the data into

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

Pros:

- Valid inference for the selected model.
- Flexibility in models: GLIMs!
- Less conservative inference than PoSI.

PoSI from Split Samples

Very different “obvious” approach: Split the data into

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

Pros:

- Valid inference for the selected model.
- Flexibility in models: GLIMs!
- Less conservative inference than PoSI.

Cons:

- Artificial randomness from a single split.
- Reduced effective sample size.
- More model selection uncertainty.
- More estimation uncertainty.
- Loss of conditionality on **X**.

Fixed $\mathbf{X}$ versus Random $\mathbf{X}$

Fixed **X** versus Random **X**

- With split-sampling we break the fixed-**X** paradigm.
- Why do many statisticians believe in conditioning on X?

Fixed **X** versus Random **X**

- With split-sampling we break the fixed-**X** paradigm.
- Why do many statisticians believe in conditioning on X?
  Answer: Fisher's ancillarity argument for **X**.

Fixed **X** versus Random **X**

- With split-sampling we break the fixed-**X** paradigm.
- Why do many statisticians believe in conditioning on X?
  Answer: Fisher's ancillarity argument for **X**.
- Scenario: $Y =$ error-free but nonlinear response
  $X =$ random predictor
  $\Rightarrow \quad Y|X$ has no randomness for fixed X

  Demo: Execute the following line in R.

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

Fixed **X** versus Random **X**

- With split-sampling we break the fixed-**X** paradigm.
- Why do many statisticians believe in conditioning on X?
  Answer: Fisher's ancillarity argument for **X**.
- Scenario: $Y =$ error-free but nonlinear response
  $X =$ random predictor
  $\Rightarrow \quad Y|X$ has no randomness for fixed X

  Demo: Execute the following line in R.

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

Nonlinearity of Y and randomness of X conspire to create sampling variability in the estimates.

Fixed **X** versus Random **X**

- With split-sampling we break the fixed-**X** paradigm.
- Why do many statisticians believe in conditioning on X?
  Answer: Fisher's ancillarity argument for **X**.
- Scenario: $Y =$ error-free but nonlinear response
  $X =$ random predictor
  $\Rightarrow$ $Y|X$ has no randomness for fixed X

  Demo: Execute the following line in R.

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

Nonlinearity of Y and randomness of X conspire to create sampling variability in the estimates.

- Consequence: Ancillarity of X is invalid if the model is an approximation.

Inference for Random $\mathbf{X}$

Inference for Random **X**

- Fact: Econometricians do not condition on **X**.
  They use an alternative form of inference based on the

  Sandwich Estimate of Standard Error.

  Eicker-**Huber**-White

  PJH, "THE BEHAVIOR OF MAXIMUM LIKELIHOOD ESTIMATES UNDER NONSTANDARD CONDITIONS", Berkeley Symp. 1967

Inference for Random **X**

- Fact: Econometricians do not condition on **X**.
  They use an alternative form of inference based on the
  Sandwich Estimate of Standard Error.
  Eicker-**Huber**-White
  PJH, "THE BEHAVIOR OF MAXIMUM LIKELIHOOD ESTIMATES UNDER NONSTANDARD CONDITIONS", Berkeley Symp. 1967
- Do statisticians know regression inference that is not conditional on **X**?

Inference for Random **X**

- Fact: Econometricians do not condition on **X**.
  They use an alternative form of inference based on the
  Sandwich Estimate of Standard Error.
  Eicker-**Huber**-White
  PJH, "THE BEHAVIOR OF MAXIMUM LIKELIHOOD ESTIMATES UNDER NONSTANDARD CONDITIONS", Berkeley Symp. 1967
- Do statisticians know regression inference that is not conditional on **X**?
  Yes, we do: the Pairs Bootstrap

Inference for Random **X**

- Fact: Econometricians do not condition on **X**.
  They use an alternative form of inference based on the
  Sandwich Estimate of Standard Error.
  Eicker-**Huber**-White
  PJH, "THE BEHAVIOR OF MAXIMUM LIKELIHOOD ESTIMATES UNDER NONSTANDARD CONDITIONS", Berkeley Symp. 1967

- Do statisticians know regression inference that is not conditional on **X**?
  Yes, we do: the Pairs Bootstrap
  to be distinguished from the Residual Bootstrap (which is fixed-**X**).

Inference for Random **X**

- Fact: Econometricians do not condition on **X**.
  They use an alternative form of inference based on the
  Sandwich Estimate of Standard Error.
  Eicker-**Huber**-White
  PJH, "THE BEHAVIOR OF MAXIMUM LIKELIHOOD ESTIMATES UNDER NONSTANDARD CONDITIONS", Berkeley Symp. 1967
- Do statisticians know regression inference that is not conditional on **X**?
  Yes, we do: the Pairs Bootstrap
  to be distinguished from the Residual Bootstrap (which is fixed-**X**).
- Fact: The Sandwich estimate of Standard Error is the limit of the
  M-of-N bootstrap as $M \to \infty$.

The Pairs Bootstrap for Regression

- Assumptions: $(\mathbf{x}_i, y_i) \sim P(d\mathbf{x}, dy)$ i.i.d.,

  $P(d\mathbf{x})$ non-degenerate: $\mathbf{E}[\mathbf{x}\mathbf{x}'] > \mathbf{0}$, + technicalities for CLTs of estimates.

- There is no regression model, but we apply regression anyway, OLS, say:

  $$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- The nonparametric pairs bootstrap applies:

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

  Note: Militant conditionalists would reject this; they would bootstrap residuals.

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

The Pairs Bootstrap for Regression

- Assumptions: $(\mathbf{x}_i, y_i) \sim P(d\mathbf{x}, dy)$ i.i.d.,

  $P(d\mathbf{x})$ non-degenerate: $\mathbf{E}[\mathbf{x}\mathbf{x}'] > \mathbf{0}$, + technicalities for CLTs of estimates.

- There is no regression model, but we apply regression anyway, OLS, say:

  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

- The nonparametric pairs bootstrap applies:

  Resample $(\mathbf{x}_i, y_i)$ pairs $\rightarrow$ $(\mathbf{x}_i^*, y_i^*)$ $\rightarrow$ $\hat{\beta}^*$.

  Note: Militant conditionalists would reject this; they would bootstrap residuals.

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

Question: Letting $\hat{\mathrm{SE}}_{\mathrm{lin}}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\|\mathbf{x}_{j\bullet}\|}$, is the following always true?

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

Conventional vs Bootstrap Std Errors: Can they differ?

Compare conventional and bootstrap standard errors:

- Boston Housing Data (no groans, please! Caveat...)
- Response: `MEDV` of single residences in a census tract, $N = 506$
- $R^2 \approx 0.74$, residual dfs $= 487$

Conventional vs Bootstrap Std Errors: Can they differ?

Compare conventional and bootstrap standard errors:

- Boston Housing Data (no groans, please! Caveat...)
- Response: MEDV of single residences in a census tract, $N = 506$
- $R^2 \approx 0.74$, residual dfs $= 487$

	$\hat{\beta}_j$	SE_{lin}	SE_{boot}	SE_{boot}/SE_{lin}	t_{lin}
CRIM	-0.099	0.031	0.033	1.074	-3.261
ZN	0.121	0.035	0.035	1.004	3.508
INDUS	0.017	0.046	0.038	0.843	0.382
CHAS	0.074	0.024	0.036	**1.503**	3.152
NOX	-0.224	0.048	0.048	1.003	-4.687
RM	0.290	0.032	0.065	**2.049**	9.149
AGE	0.002	0.040	0.050	1.236	0.044
DIS	-0.344	0.045	0.048	1.068	-7.598
RAD	0.288	0.062	0.060	0.958	4.620
TAX	-0.233	0.068	0.051	**0.740**	-3.409
PTRATIO	-0.218	0.031	0.026	0.865	-7.126
B	0.092	0.026	0.027	1.036	3.467
LSTAT	-0.413	0.039	0.078	**1.995**	-10.558

Conventional vs Bootstrap Std Errors (contd.)

Compare conventional and bootstrap standard errors:

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

Conventional vs Bootstrap Std Errors (contd.)

Compare conventional and bootstrap standard 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

Conventional vs Bootstrap Std Errors (contd.)

Compare conventional and bootstrap standard errors:

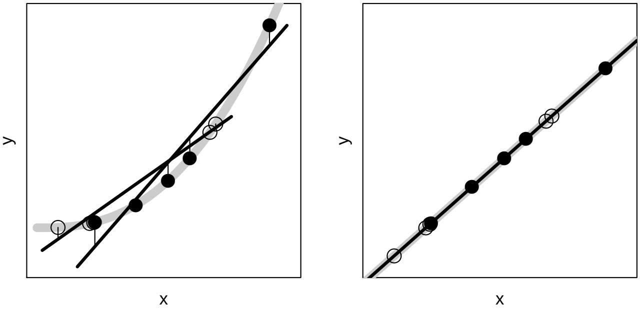
- LA Homeless Data (Richard Berk, UPenn)
- Response: `StreetTotal` of homeless in a census tract, $N = 505$
- $R^2 \approx 0.13$, residual dfs $= 498$

	$\hat{\beta}_j$	$\mathrm{SE}_{\mathrm{lin}}$	$\mathrm{SE}_{\mathrm{boot}}$	$\mathrm{SE}_{\mathrm{boot}}/\mathrm{SE}_{\mathrm{lin}}$	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

- Which standard errors are we to believe?
- What is the reason for the discrepancy?
- Is the pairs bootstrap a failure?

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

Recall the demo: A noise-free nonlinearity $y_i = \mu(\mathbf{x}_i) \sim x_i^2, \quad x_i$ i.i.d. fitted by a straight line.

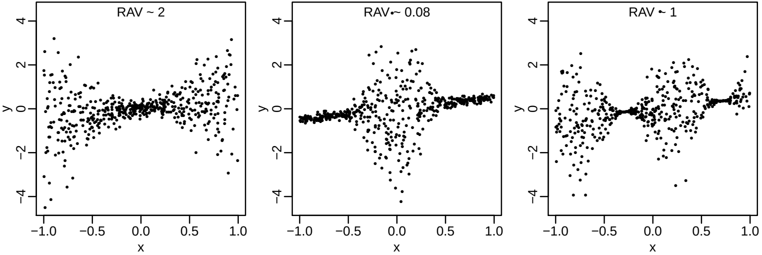


Nonlinearity + randomness of X = sampling variability.

- Hal White[†2012] (1980), "Using Least Squares to Approximate Unknown Regression Functions," International Economic Review

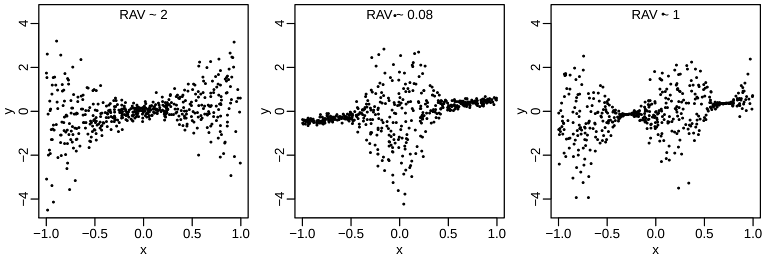
Second Reason for $\mathrm{SE_{boot}} \neq \mathrm{SE_{lin}}$: Heteroskedasticity

Which has the smallest/largest true $\mathrm{SE}(\hat{\beta})$? ($\sum \sigma_i^2$ are the same.)



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

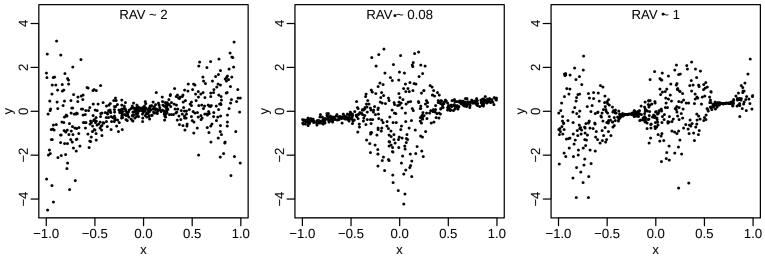
Which has the smallest/largest true $\mathrm{SE}(\hat{\beta})$? ($\sum \sigma_i^2$ are the same.)



SE=0.10 SE=0.02 SE=0.07

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

Which has the smallest/largest true $\mathrm{SE}(\hat{\beta})$? ($\sum \sigma_i^2$ are the same.)



SE=0.10 SE=0.02 SE=0.07

Heteroskedasticity can invalidate Linear Model SEs.

- Hinkley (1977) “Jackknifing in Unbalanced Situations,” Technometrics
- Wu (1986) “Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis,” AoS
- Hal White$^{\dagger 2012}$ (1980), “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” Econometrica (1980)

But Why Would Anyone Use an "Incorrect" Model?

- Often we don't know that the model is violated by the data.
  $\Longrightarrow$ An argument in favor of diligent model diagnostics...

- The problem persists even if we use basis expansion but miss the nature of the nonlinearity: curves, jaggies, jumps, ...

- Linear models provide low-df approximations which may be all that is feasible when p is large compared to n.

- Even when the model is only an approximation, the slopes contain information about the direction of the association.

- $\exists$ interpretations of slopes w/o assuming a correct model:

weighted averages of "case slopes"

$$\hat{\beta} = \sum_{i=1\ldots n} w_i \hat{\beta}_i, \qquad \hat{\beta}_i = \frac{y_i - \bar{y}}{x_i - \bar{x}}, \qquad w_i = \frac{(x_i - \bar{x})^2}{\sum_{k=1..n}(x_k - \bar{x})^2}.$$

Redefining the Population and the Parameters

- Joint distribution, i.i.d. sampling: $(\mathbf{x}_i, y_i) \sim P(d\mathbf{x}, dy)$

  Assume properties sufficient to grant CLTs for estimates of interest.

- No assumptions on $\mu(\mathbf{x}) = \mathbf{E}[\, y \,|\, \mathbf{x}\,]$, $\sigma^2(\mathbf{x}) = \mathbf{V}[\, y \,|\, \mathbf{x}\,]$.

- Define a population OLS parameter:

$$\beta \;:=\; \mathrm{argmin}_{\tilde{\beta}}\, \mathbf{E}\left[\left(y - \tilde{\beta}'\mathbf{x}\right)^2\right] \;=\; \mathbf{E}[\,\mathbf{x}\,\mathbf{x}'\,]^{-1}\,\mathbf{E}[\,\mathbf{x}\,y\,]$$

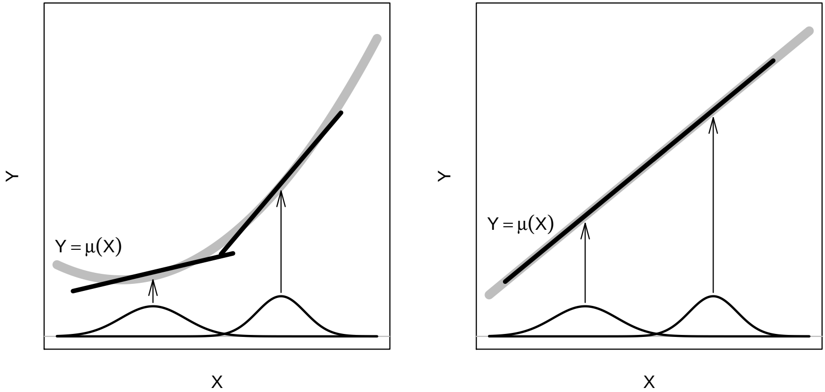
- This is the target of inference: $\beta \;=\; \beta(P)$

  Thus β is a statistical functional, not a generative parameter.

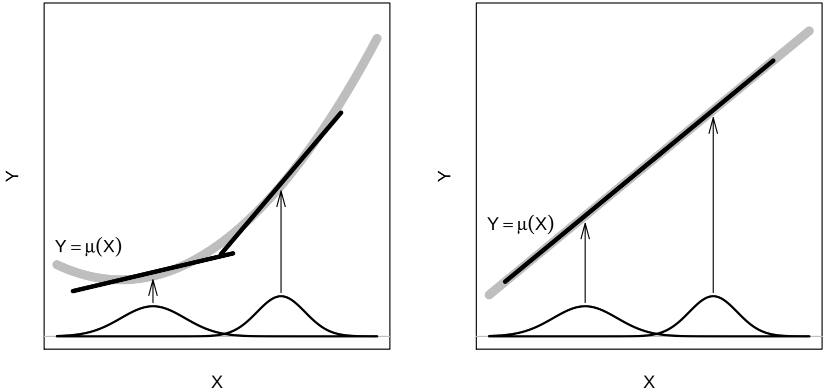
$\Longrightarrow$ **"Statistical Functional View of OLS"**

("Random $\mathbf{X}$ Theory")

The LS Population Parameter



The LS Population Parameter



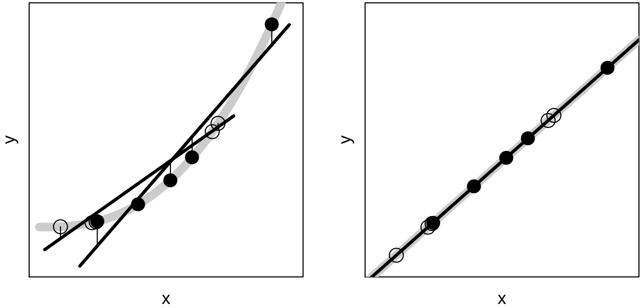
If $\mu(\mathbf{x})$ is nonlinear, $\beta(\mathbf{P})$ depends on the x-distribution $\mathbf{P}(d\mathbf{x})$!

The LS Estimator and its Target

- Data: $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)'$, $\quad \mathbf{y} = (y_1, ..., y_N)'$,
- Target of estimation and inference in linear models theory:

$$\beta(\mathbf{X}) \;=\; \mathbf{E}[\hat{\beta}|\mathbf{X}] \;=\; (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{E}[\mathbf{y}|\mathbf{X}]$$

- When $\mu(\mathbf{x}) = \mathbf{E}[y|\mathbf{x}]$ is nonlinear, then $\beta(\mathbf{X})$ is a random vector.



Linear Models Theory versus Econometrics

- Consider the simplest case of a single predictor, no intercept, and define the conditional MSE by $m^2(x) := \mathbf{E}[(Y - \beta' x)^2 | x]$
- The **correct** asymptotic variance of $\hat{\beta}$ is

$$\boldsymbol{AV}_{sand} = \frac{\mathbf{E}[m^2(x)x^2]}{\mathbf{E}[x^2]^2}.$$

- If we were to use standard errors from linear models theory, the following **incorrect** asymptotic variance is implied:

$$\boldsymbol{AV}_{lin} = \frac{\mathbf{E}[m^2(\mathbf{x})]}{\mathbf{E}[x^2]}$$

- Define the "**Ratio of Asymptotic Variances**" or **RAV**:

$$\mathbf{RAV} := \frac{\boldsymbol{AV}_{sand}}{\boldsymbol{AV}_{lin}} = \frac{\mathbf{E}[m^2(x)x^2]}{\mathbf{E}[m^2(\mathbf{x})]\,\mathbf{E}[x^2]}$$

Linear Models Theory versus Econometrics (contd.)

$$\mathbf{RAV} = \frac{\boldsymbol{AV}_{sand}}{\boldsymbol{AV}_{lin}} = \frac{\mathbf{E}[m^2(\mathbf{x})x^2]}{\mathbf{E}[m^2(x)]\mathbf{E}[x^2]}$$

Linear Models Theory versus Econometrics (contd.)

$$\mathbf{RAV} \;=\; \frac{\mathbf{AV}_{sand}}{\mathbf{AV}_{lin}} \;=\; \frac{\mathbf{E}[m^2(\mathbf{x})x^2]}{\mathbf{E}[m^2(x)]\mathbf{E}[x^2]}$$

- **Fact:**

$$\max_m \mathbf{RAV} \;=\; \infty, \qquad \min_m \mathbf{RAV} \;=\; 0$$

- Conclusion: Asymptotically the discrepancy between SE_{sand} and SE_{lin} can be arbitrarily large in either direction.
- In practice, **RAV** $>$ 1 is more frequent **and** more dangerous because it invalidates conventional linear models inference.

Next Steps, in Outline only

Next Steps, in Outline only

- The discrepancy between SE_{lin} and SE_{sand} can be turned into a diagnostic test.

Next Steps, in Outline only

- The discrepancy between SE_{lin} and SE_{sand} can be turned into a diagnostic test.
- A robustness problem:
  - Asymptotic variance is a 4^{th} order functional.
    $\Rightarrow \mathrm{SE}_{sand}$ is even less robust than SE_{lin}.
  - The robustness problem is equally present in SE_{sand} and SE_{boot}.

Next Steps, in Outline only

- The discrepancy between SE_{lin} and SE_{sand} can be turned into a diagnostic test.
- A robustness problem:
  - Asymptotic variance is a 4^{th} order functional.
    $\Rightarrow \mathrm{SE}_{sand}$ is even less robust than SE_{lin}.
  - The robustness problem is equally present in SE_{sand} and SE_{boot}.
- A new PoSI technology can be based on asymptotic normality and estimates of **AV**:
  - Sandwich/bootstrap PoSI computations become slightly more expensive: Initial reduction is to $(p+1)p/2$ rather than p dimensions.
  - Sandwich/bootstrap PoSI allows us to protect against selection of a finite dictionary of transformations in addition to selection of predictors. $(g(Y), f_1(X_1), ..., f_p(X_p))$ is no different than $(Y, X_1, ..., X_p)$.

Some Take-Home Points about Approximate Models

- Robustness should include not just misspecification of error distributions but of 1st and 2nd order misspecifications as well.
  $\Rightarrow$ Sandwich or pairs-bootstrap estimates of standard error
- Beware of the ancillarity fallacy: Ancillarity arguments are invalidated by 1st order model misspecifications.
- Fixed-**X** standard errors SE_{lin} can be substantial underestimates of true sampling variation; the opposite can occur, too, but less often.
- In any regression, not all predictors are equally affected by standard error discrepancies.

Back to the Big Picture: Reproducibility

- Contributing factor to non-reproducibility:

  Unaccounted data-analytic activities such as

  - selection of predictor variables*
  - selection of outcome variables*
  - selection of data transformations**
  - informal EDA before formal model selection*
  - informal diagnostics after formal model selection*
  - meta-selection of selection methods*

  * solved by fixed-**X** PoSI under 1st order misspecific. & homoskedasticity.

- From a fixed-**X** to a random-**X** framework:
  - Correct inference under minimal assumptions: $(y_i, \mathbf{x}_i) \sim$ iid
  - Accounts for nonlinearity and heteroskedasticity.
  - Permits PoSI for selection of transformations; solves **.

THANKS