Post-Selection Inference for Models that are Approximations

Andreas Buja

joint work 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

2013/11/13

Larger Problem: Non-Reproducible Empirical Findings

Larger Problem: Non-Reproducible Empirical Findings

- Indicators of a problem

  (from: Berger, 2012, “Reproducibility of Science: P-values and Multiplicity”)

  - Bayer Healthcare reviewed 67 in-house attempts at replicating findings in published research:

    $< 1/4$ were viewed as replicated.
  - Arrowsmith (2011, Nat. Rev. Drug Discovery 10):

    Increasing failure rate in Phase II drug trials
  - Ioannidis (2005, PLOS Medicine):

    “Why Most Published Research Findings Are False”

Larger Problem: Non-Reproducible Empirical Findings

- Indicators of a problem

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

  - Bayer Healthcare reviewed 67 in-house attempts at replicating findings in published research:

    $< 1/4$ were viewed as replicated.
  - Arrowsmith (2011, Nat. Rev. Drug Discovery 10):

    Increasing failure rate in Phase II drug trials
  - Ioannidis (2005, PLOS Medicine):

    "Why Most Published Research Findings Are False"

- Many potential causes:
  - publication biases
  - economic biases
  - experimental biases
  - statistical biases
  - ...

Statistical Biases – one among several

- Hypothesis: A statistical bias is due to

  an absence of accounting for model/variable selection.

Statistical Biases – one among several

- Hypothesis: A statistical bias is due to

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: AIC, BIC, Lasso, ...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: “The effect size is too small in relation to the cost of data collection to warrant inclusion of this predictor.”

Statistical Biases – one among several

- Hypothesis: A statistical bias is due to

  an absence of accounting for model/variable selection.

- Model selection is done on several levels:
  - formal selection: AIC, BIC, Lasso, ...
  - informal selection: residual plots, influence diagnostics, ...
  - post hoc selection: "The effect size is too small in relation to the cost of data collection to warrant inclusion of this predictor."
- 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.
  - If we develop valid post-selection inference for "adaptive Lasso", say, it won't solve the problem because few empirical researchers would 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

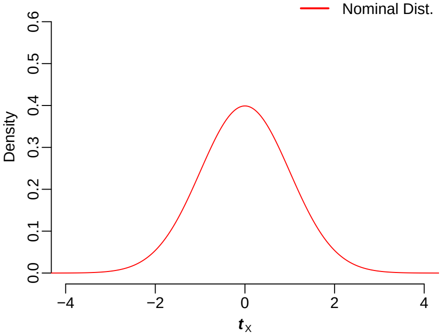
In common practice:

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

Is this inference valid?

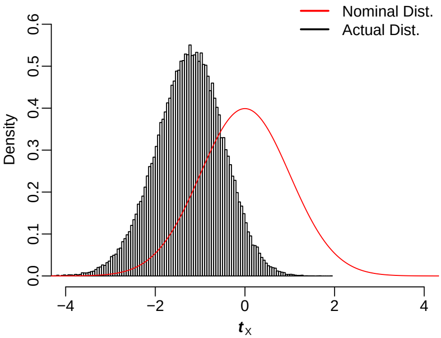
Evidence from a Simulation

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!

- But first we need some preliminaries on

Targets of Inference and Inference in **Approximate** Models

Submodels — Notation, Parameters, Assumptions

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

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

- The LS estimators in the submodel M are

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

- What does $\hat{\beta}_{\mathrm{M}}$ estimate, **not** assuming the truth of M?
  A: Its expectation — i.e., we ask for unbiasedness.

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

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

- Once again: We do not assume that the submodel is correct,
  i.e., we allow $\mu \neq \mathbf{X}_{\mathrm{M}} \beta_{\mathrm{M}}$! But $\mathbf{X}_{\mathrm{M}} \beta_{\mathrm{M}}$ is the best approximation to μ.

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:

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

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

$$t_{j\bullet\mathrm{M}} \;:=\; \frac{\hat{\beta}_{j\bullet\mathrm{M}} - \beta_{j\bullet\mathrm{M}}}{\hat{\sigma}/\|\mathbf{X}_{j\bullet\mathrm{M}}\|} \;=\; \frac{\mathbf{X}_{j\bullet\mathrm{M}}^T(\mathbf{Y}-\mathbf{E}[\mathbf{Y}])}{\|\mathbf{X}_{j\bullet\mathrm{M}}\|\;\hat{\sigma}}.$$

Parameters One More Time

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

  $\boxed{\mathrm{M} \neq \mathrm{M}' \;\Rightarrow\; \beta_{j \bullet \mathrm{M}} \neq \beta_{j \bullet \mathrm{M}'}}$ in value and in meaning.

  Motto: A difference in adjustment implies a difference in parameters.

- It follows that there are up to $p\,2^{p-1}$ different parameters $\beta_{j \bullet \mathrm{M}}$!!!

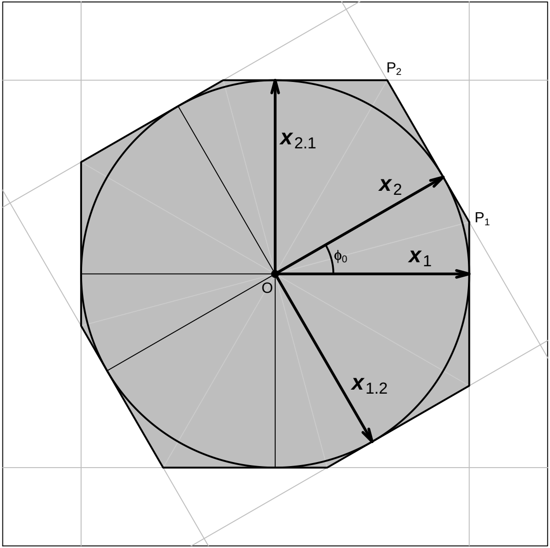
- However, they are intrinsically p-dimensional:

$$\boldsymbol{\beta}_{\mathrm{M}} = (\mathbf{X}_{\mathrm{M}}^{T}\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^{T}\,\mathbf{X}\,\boldsymbol{\beta}$$

  where $\mathbf{X}$ and $\boldsymbol{\beta}$ are from the full model.

- Hence each $\beta_{j \bullet \mathrm{M}}$ is a lin. comb. of the full model parameters $\beta_1, ..., \beta_p$.

Geometry of Adjustment



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

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}}\hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M; (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model.

$\Longrightarrow$ $t_{j\bullet\mathrm{M}}$ will have a t-distribution with the same dfs $\forall\mathrm{M}$, $\forall j \in \mathrm{M}$.

Error Estimates $\hat{\sigma}^2$

- Important: To enable simultaneous inference for all $t_{j\bullet\mathrm{M}}$,
  - do not use the error estimate $\cancel{\hat{\sigma}^2_{\mathrm{M}}} := \|\mathbf{Y} - \mathbf{X}_{\mathrm{M}}\hat{\boldsymbol{\beta}}_{\mathrm{M}}\|^2/(n-m)$ in M; (the selected model M may well be 1st order wrong;)
  - instead, for all models M use $\hat{\sigma}^2 = \hat{\sigma}^2_{Full}$ from the full model.

$\Longrightarrow$ $t_{j\bullet\mathrm{M}}$ will have a t-distribution with the same dfs $\forall\mathrm{M},\ \forall j \in \mathrm{M}$.

- What if even the full model is 1st order wrong?
  Answer: $\hat{\sigma}^2_{Full}$ will be inflated and inference will be conservative.
  But better estimates are available if ...
  - exact replicates exist: use $\hat{\sigma}^2$ from the 1-way ANOVA of replicates;
  - a larger than the full model can be assumed 1st order correct: use $\hat{\sigma}^2_{Large}$;
  - a previous dataset provided a valid estimate: use $\hat{\sigma}^2_{previous}$;
  - nonparametric estimates are available: use $\hat{\sigma}^2_{nonpar}$ (Hall and Carroll 1989).

PS: In the fashionable $p > N$ literature, what is their $\hat{\sigma}^2$?

Statistical Inference under First Order Incorrectness

- Statistical inference, one parameter at a time:
  If $r =$ dfs in $\hat{\sigma}^2$ and $K = t_{1-\alpha/2,r}$, then the confidence intervals

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

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

Statistical Inference under First Order Incorrectness

- Statistical inference, one parameter at a time:
  If $r =$ dfs in $\hat{\sigma}^2$ and $K = t_{1-\alpha/2,r}$, then the confidence intervals

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

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

- Achieved so far:

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

  - No assumption is made that the submodels are 1st order correct;
  - Even the full model may be 1st order incorrect
    if a valid $\hat{\sigma}^2$ is otherwise available.
  - A single error estimate opens up the possibility of
    simultaneous inference across submodels.

Variable Selection

- What is a variable selection procedure?

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

  - $\hat{\mathrm{M}}$ 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**, ...).

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**, ...).

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

  - $\hat{\mathrm{M}}$ 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**, ...).

- Candidate for meaningful coverage probabilities:

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

- Problem: No such coverage probabilities are known or can be estimated for most selection procedures $\hat{\mathrm{M}}$.

Variable Selection

- What is a variable selection procedure?

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

  - $\hat{\mathrm{M}}$ divides the response space $\mathrm{I\!R}^N$ 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**, ...).

- Candidate for meaningful coverage probabilities:

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

- Problem: No such coverage probabilities are known or can be estimated for most selection procedures $\hat{\mathrm{M}}$.

- Solution: Ask for more! It is possible to construct 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":*

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

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":

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

Theorem

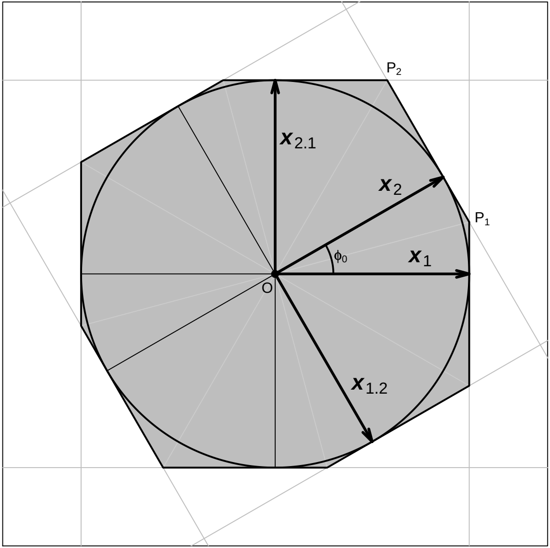
Let K be the $1-\alpha$ quantile of the "**max-max-**$|t|$" *statistic of the lemma:*

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

Then we have the following universal PoSI guarantee:

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

PoSI Geometry — Simultaneity



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

Computing PoSI

- The simultaneity challenge: there are $p\,2^{p-1}$ statistics $|t_{j\bullet\mathrm{M}}|$.

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

- Monte Carlo-approximation in R, brute force, up to $p \approx 20$.
- Computations are specific to a design $\mathbf{X}$: $K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(\mathbf{X}, \alpha, df)$
- One Monte Carlo computation is good for any α and any error df.

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \; \hat{\sigma}} \;\sim\; \sqrt{p\, F_{p,df}}.$$

  - The Scheffé method provides sim. inference for all linear "contrasts".
  - The Scheffé constant is $K_{\mathrm{Sch}} = K_{\mathrm{Sch}}(p, \alpha, df) = \sqrt{p\, F_{p,df;1-\alpha}}$.

Scheffé Protection Yields Valid PoSI

- **Scheffé Simultaneous Inference** is based on the statistic

$$\sup_{\mathbf{x} \in \mathrm{col}(\mathbf{X}) \setminus \{\mathbf{0}\}} \frac{|\mathbf{x}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{x}\| \, \hat{\sigma}} \sim \sqrt{p \, F_{p,df}}.$$

  - The Scheffé method provides sim. inference for all linear "contrasts".
  - The Scheffé constant is $K_{\mathrm{Sch}} = K_{\mathrm{Sch}}(p, \alpha, df) = \sqrt{p \, F_{p,df;1-\alpha}}$.

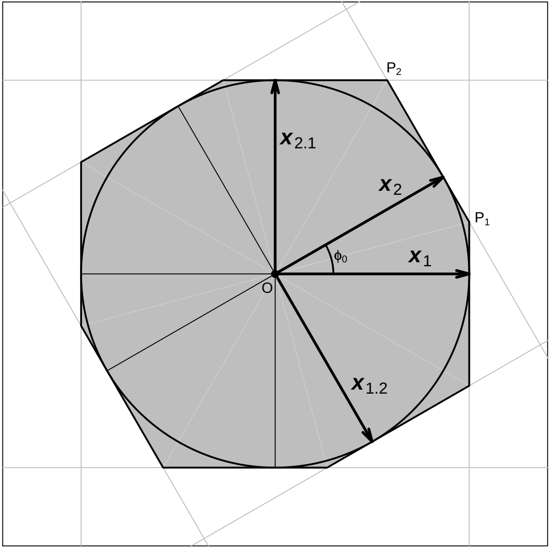
- Compare: **PoSI Simultaneous Inference** is based on the statistic

$$\max_{\mathrm{M}} \max_{j \in \mathrm{M}} \frac{|\mathbf{X}_{j \bullet \mathrm{M}}^T(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])|}{\|\mathbf{X}_{j \bullet \mathrm{M}}\| \, \hat{\sigma}}$$

  The PoSI contrasts are a subset of the Scheffé contrasts, hence:

  - Scheffé statistic $\geq$ PoSI statistic
  - $K_{\mathrm{Sch}} \geq K_{\mathrm{PoSI}}$
  - Scheffé yields universally valid conservative PoSI.

The Scheffé Ball and the PoSI Polytope



Circle = Scheffé Ball

The PoSI polytope is tangent to the ball.

PoSI Benefits

PoSI protection may be very conservative, but it has benefits:

- One can try many selection methods and pick the “best” by whatever standard. The PoSI-significant coefficents will be valid.
- One can perform informal model diagnostics and change one’s mind based on them, PoSI inference will still be valid.
- After computing PoSI, one can go on fishing expeditions among models and search for significances based on PoSI. The fishing will not invalidate the inference.
- In a clinical trial, one can perform post-hoc “data mining” for significant effects, and the PoSI-protected findings will be valid.

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**.

Conditionality on **X**: Fixed versus Random **X**

- With split-sampling we have broken conditionality on **X**: random splitting means the predictors are treated as random.
- Why do some statisticians insist on fixed-**X** regression?

Conditionality on **X**: Fixed versus Random **X**

- With split-sampling we have broken conditionality on **X**: random splitting means the predictors are treated as random.
- Why do some statisticians insist on fixed-**X** regression? Answer: Fisher's ancillarity argument for **X**

Conditionality on **X**: Fixed versus Random **X**

- With split-sampling we have broken conditionality on **X**:
  random splitting means the predictors are treated as random.
- Why do some statisticians insist on fixed-**X** regression?
  Answer: Fisher's ancillarity argument for **X**
- Fact: Econometricians do not condition on **X**.
  They use an alternative form of inference based on the
  Sandwich Estimate of Standard Error.

Conditionality on $\mathbf{X}$: Fixed versus Random $\mathbf{X}$

- With split-sampling we have broken conditionality on $\mathbf{X}$: random splitting means the predictors are treated as random.
- Why do some statisticians insist on fixed-$\mathbf{X}$ regression? Answer: Fisher's ancillarity argument for $\mathbf{X}$
- Fact: Econometricians do not condition on $\mathbf{X}$. They use an alternative form of inference based on the Sandwich Estimate of Standard Error.
- Do we know regression inference that is not conditional on $\mathbf{X}$?

Conditionality on **X**: Fixed versus Random **X**

- With split-sampling we have broken conditionality on **X**:
  random splitting means the predictors are treated as random.
- Why do some statisticians insist on fixed-**X** regression?
  Answer: Fisher's ancillarity argument for **X**
- Fact: Econometricians do not condition on **X**.
  They use an alternative form of inference based on the
  Sandwich Estimate of Standard Error.
- Do we 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**).

The Pairs Bootstrap for Regression

The Pairs Bootstrap for Regression

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

  P 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, LS, say:

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

- The nonparametric paired 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 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, LS, say:

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

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

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 $= 487$

	$\hat{\beta}_j$	$\mathrm{SE}_{\mathrm{lin}}$	$\mathrm{SE}_{\mathrm{boot}}$	$\mathrm{SE}_{\mathrm{boot}}/\mathrm{SE}_{\mathrm{lin}}$	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 paired bootstrap a failure?

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

- Consider a noise-free nonlinearity,

$$y_i \;=\; \mu(\mathbf{x}_i) \;\sim\; x_i^2, \quad x_i \ \text{i.i.d.}$$

and fit a straight line anyway. Watch the effect:

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

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

- Consider a noise-free nonlinearity,

$$y_i \;=\; \mu(\mathbf{x}_i) \;\sim\; x_i^2, \quad x_i \;\text{i.i.d.}$$

and fit a straight line anyway. Watch the effect:

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

Nonlinearity and **randomness of X** conspire
to create sampling variability.

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

- Consider a noise-free nonlinearity,

$$y_i \;=\; \mu(\mathbf{x}_i) \;\sim\; x_i^2, \quad x_i \text{ i.i.d.}$$

and fit a straight line anyway. Watch the effect:

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

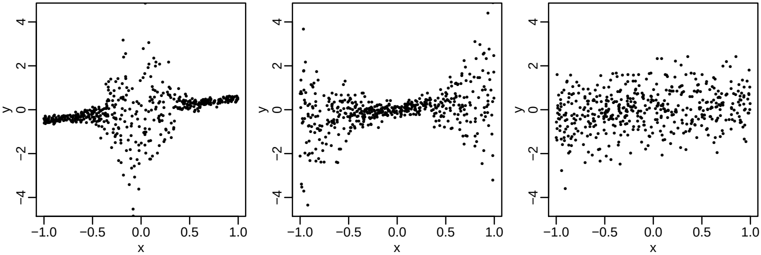
Nonlinearity and **randomness of X** conspire
to create sampling variability.

- “Econometrics 101”: Hal White$^{\dagger 2012}$ (1980)

“Using Least Squares to Approximate Unknown Regression Functions,” Intl. Economic Review.

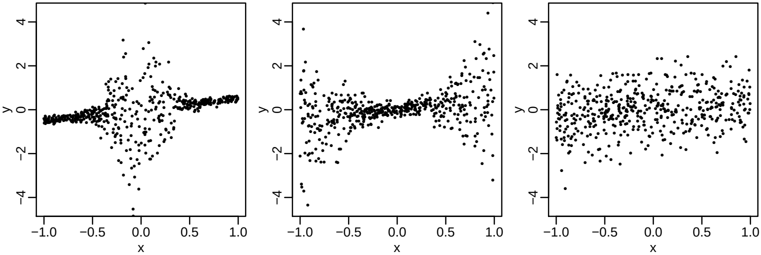
Second Reason for $\mathrm{SE}_{\mathrm{boot}} \neq \mathrm{SE}_{\mathrm{lin}}$

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



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

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

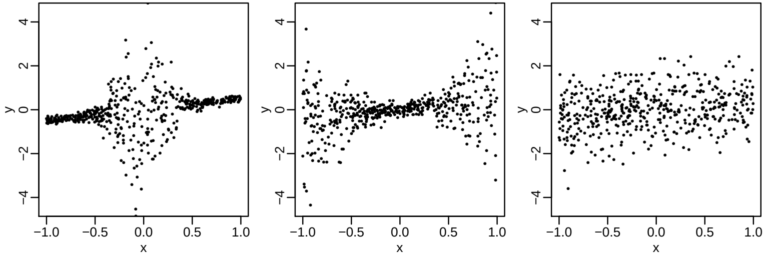


$\mathrm{SE}{=}0.02$ $\mathrm{SE}{=}0.10$ $\mathrm{SE}{=}0.07$

- Heteroskedasticity: $\sigma^2(\mathbf{x}) := \mathbf{V}[y \,|\, \mathbf{x}] \neq \mathrm{const}$

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

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

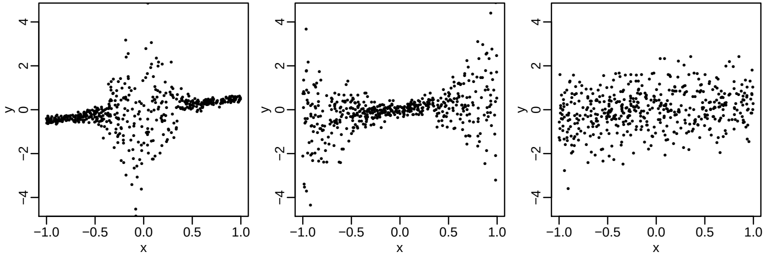


SE=0.02 SE=0.10 SE=0.07

- Heteroskedasticity: $\sigma^2(\mathbf{x}) := \mathbf{V}[y\,|\mathbf{x}] \neq \mathrm{const}$
- Tradition in statistics – conditional on $\mathbf{X}$:
  - Hinkley: “Jackknifing in Unbalanced Situations,” Technometrics (1977)
  - Wu: “Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis,” AoS (1986).

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

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



SE=0.02 SE=0.10 SE=0.07

- Heteroskedasticity: $\sigma^2(\mathbf{x}) := \mathbf{V}[y \,|\mathbf{x}] \neq \mathrm{const}$
- Tradition in statistics – conditional on $\mathbf{X}$:
  - Hinkley: "Jackknifing in Unbalanced Situations," Technometrics (1977)
  - Wu: "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," AoS (1986).
- "Econometrics 101": 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?

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...

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, ...

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 n/p is small.

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

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 n/p is small.
- 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

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.

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 $\mu(\mathbf{x}) = \mathbf{E}[\,y \mid \mathbf{x}\,], \quad \sigma^2(\mathbf{x}) = \mathbf{V}[\,y \mid \mathbf{x}\,].$

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 $\mu(\mathbf{x}) = \mathbf{E}[\, y \,|\, \mathbf{x} \,], \quad \sigma^2(\mathbf{x}) = \mathbf{V}[\, y \,|\, \mathbf{x} \,].$

- Define a population LS parameter:

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

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 LS parameter:

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

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

  Thus $\boldsymbol{\beta}$ is a statistical functional, not a generative parameter.

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 LS parameter:

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

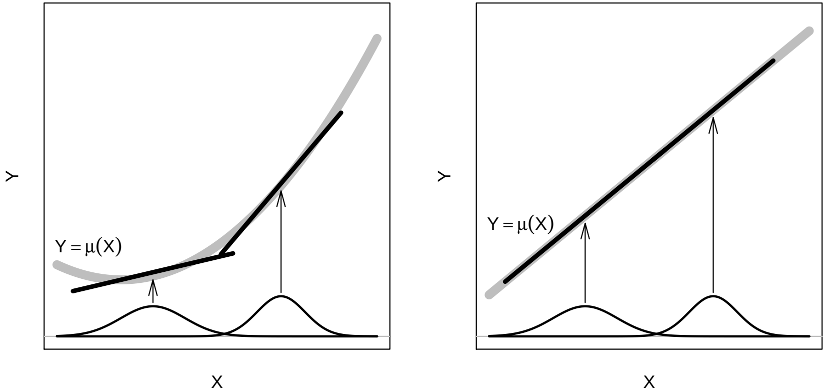
- This is the target of inference: $\boldsymbol{\beta} \;=\; \boldsymbol{\beta}(P)$

  Thus $\boldsymbol{\beta}$ is a statistical functional, not a generative parameter.

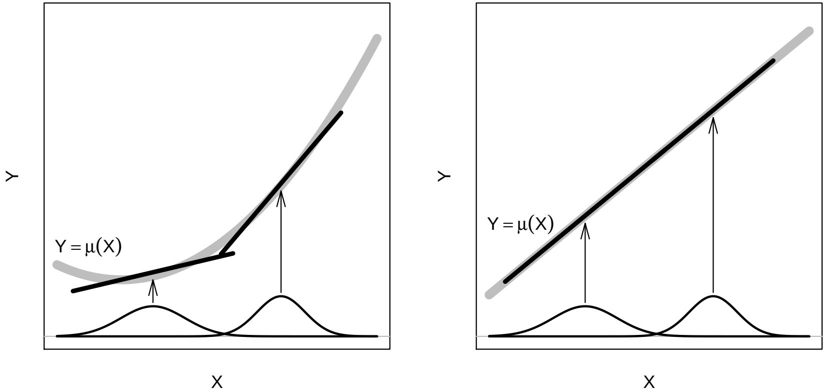
$\Longrightarrow$ **"Semi-parametric LS framework"**

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

The LS Population Parameter



The LS Population Parameter



- If $\mu(\mathbf{x})$ is nonlinear, $\boldsymbol{\beta}(\mathbf{P})$ depends on the **x**-distribution $\mathbf{P}(d\mathbf{x})$.
- If $\mu(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$ is linear, then $\boldsymbol{\beta} = \boldsymbol{\beta}(\mathbf{P})$ is the same for all $\mathbf{P}(d\mathbf{x})$).

The LS Estimator and its Target

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:

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

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.

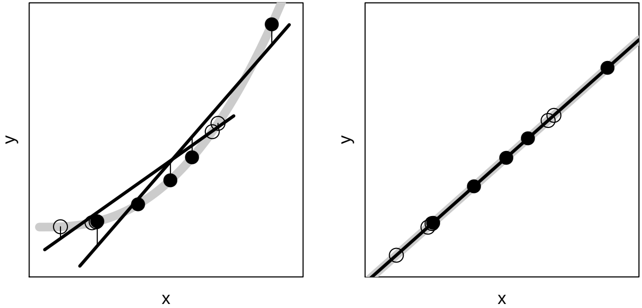
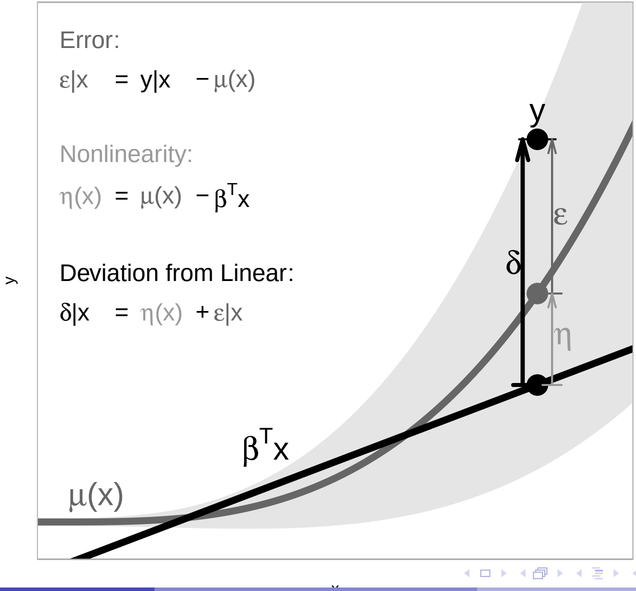


Illustration of Nonlinearity and Heteroskedasticity



The CLT for Random-**X** and the Sandwich Formula

CLT:

$$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\;\xrightarrow{\mathcal{D}}\;\; \mathcal{N}\left(\mathbf{0},\; \mathbf{E}[\,\mathbf{x}\mathbf{x}^T]^{-1}\,\mathbf{E}[\,(Y - \boldsymbol{\beta}'\mathbf{x})^2\mathbf{x}\mathbf{x}^T]\;\mathbf{E}[\,\mathbf{x}\mathbf{x}^T]^{-1}\right)$$

Simplest Sandwich Estimator of Asymptotic Variance:

$$\hat{\boldsymbol{AV}}_{sand} \;\;:=:\;\; \hat{\boldsymbol{E}}[\,\mathbf{x}\mathbf{x}^T]^{-1}\,\hat{\boldsymbol{E}}[\,(Y - \hat{\boldsymbol{\beta}}'\mathbf{x})^2\mathbf{x}\mathbf{x}^T]\;\hat{\boldsymbol{E}}[\,\mathbf{x}\mathbf{x}^T]^{-1}$$

Linear Models Theory Estimator of Asymptotic Variance:

$$\boldsymbol{AV}_{lin} \;\;:=:\;\; \hat{\boldsymbol{E}}[\,\mathbf{x}\mathbf{x}^T]^{-1}\,\hat{\sigma}^2$$

Comparing Conditional $\neq$ Unconditional Inference

- 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 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, it would mean using the following **incorrect** asymptotic variance:

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

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

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

Extremes of Conditional $\neq$ Unconditional Inference

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

Extremes of Conditional $\neq$ Unconditional Inference

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

- **Fact:**

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

Extremes of Conditional $\neq$ Unconditional Inference

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

- **Fact:**

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

- Conclusion: Asymptotically the discrepancy between conditional and unconditional SEs can be arbitrarily large.

Extremes of Conditional $\neq$ Unconditional Inference

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

- **Fact:**

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

- Conclusion: Asymptotically the discrepancy between conditional and unconditional SEs can be arbitrarily large.

- In practice, **RAV** > 1 is more frequent **and** more dangerous because it invalidates conventional linear models inference.

Outlook

- Big Benefit: The semi-parametric LS framework permits arbitrary joint $(y, \mathbf{x})$ distributions.
  - The assumption of homoskedasticity in the PoSI framework can be given up if its inference is based on sandwich or pairs bootstrp.
  - Valid inference is obtained also for arbitrarily transformed data $(f(y), g(\mathbf{x}))$.
  - If a new PoSI technology is constructed based on the semi-parametric LS framework, it will allow us to protect against selection of a finite dictionary of transformations in addition to selection of predictors.

Outlook

- Big Benefit: The semi-parametric LS framework permits arbitrary joint $(y, \mathbf{x})$ distributions.
  - The assumption of homoskedasticity in the PoSI framework can be given up if its inference is based on sandwich or pairs bootstrp.
  - Valid inference is obtained also for arbitrarily transformed data $(f(y), g(\mathbf{x}))$.
  - If a new PoSI technology is constructed based on the semi-parametric LS framework, it will allow us to protect against selection of a finite dictionary of transformations in addition to selection of predictors.
- Some obstacles:
  - Asymptotic variance is a 4^{th} order functional of the underlying distribution.
  - Hence sandwich estimates of standard error are highly non-robust. A small fraction of the data can determine the standard error estimates.
  - We may have to abandon LS and look for estimation methods whose asymptotic variances are more robust.
  - Computations of PoSI methods based on the semi-parametric framework will be even more expensive, but this should not deter us.

THANKS!