

A Probability Coverage Problem that Could Solve the Reproducibility Crisis in the Empirical Sciences

Andreas Buja

joint work with

Arun Kuchibhotla, Richard Berk, Larry Brown[†],
Ed George, Kai Zhang, Linda Zhao

Department of Statistics, The Wharton School
University of Pennsylvania
Philadelphia, USA

Flatiron 2019/03/21

Non-Reproducible Empirical Findings

- Distinction: Prediction versus Explanation
 - ▶ Prediction: estimate outcome $\hat{y}(\mathbf{x})$ from input variables $\mathbf{x} = (x_1, \dots, x_p)'$
 - ▶ Explanation: Does input variable x_j “matter” for y ?
- The Reproducibility Crisis: False Explanation
 - ▶ Ioannidis (2005, PLOS Medicine):
“Why Most Published Research Findings Are False”
 - ▶ Simmons, Nelson, Simonsohn (2011, Psychol.Sci):
“False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant”
- Many potential causes – two major ones:
 - ▶ Publication bias: “file drawer problem” (Rosenthal 1979)
 - ▶ Statistical biases: “researcher degrees of freedom” (SNS 2011),
 p -hacking, “data dredging”, “reading tea leaves” with data visualization, ...

Data Dredging: Variable Selection in Regression

- Example: Length of Criminal Sentence
- Question: What covariates predict length of a criminal sentence best?
- A small empirical study:
 - ▶ $N = 250$ observations.
 - ▶ Response: log-length of sentences
 - ▶ $p = 11$ covariates (regressors, predictors, explanatory variables):
 - ★ race
 - ★ gender
 - ★ initial age
 - ★ marital status
 - ★ employment status
 - ★ seriousness of crime
 - ★ psychological problems
 - ★ education
 - ★ drug related
 - ★ alcohol usage
 - ★ prior record
- What variables should be included?

Data Dredging: Length of Criminal Sentence (contd.)

- All-subset search with BIC chooses a model $\hat{M}$ with seven variables:

- ▶ initial age
- ▶ gender
- ▶ employment status
- ▶ seriousness of crime
- ▶ drugs related
- ▶ alcohol usage
- ▶ prior records

- t -statistics of selected covariates, in descending order:

- ▶ $|t_{\text{alcohol}}| = 3.95;$
- ▶ $|t_{\text{prior records}}| = 3.59;$
- ▶ $|t_{\text{seriousness}}| = 3.57;$
- ▶ $|t_{\text{drugs}}| = 3.31;$
- ▶ $|t_{\text{employment}}| = 3.04;$
- ▶ $|t_{\text{initial age}}| = 2.56;$
- ▶ $|t_{\text{gender}}| = 2.33.$

- Which t -values are credible evidence?

Linear Models and OLS Estimation

- Linear Model: $y = \mathbf{x}'\boldsymbol{\beta} + \epsilon$ (notation!)
- Observations: $(\mathbf{x}_i, y_i), i = 1, \dots, N, \quad \mathbf{x}_i \in \mathbb{R}^p$
 $\mathbf{Y} = (y_1, \dots, y_N)' \in \mathbb{R}^N,$
 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)' = (\mathbf{X}_1, \dots, \mathbf{X}_p) \in \mathbb{R}^{N \times p}$
 N rows p columns
- Linear Model, vectorized: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \mathbf{I}_N), \quad \boldsymbol{\beta} \in \mathbb{R}^p$
- OLS Estimate of coefficient vector:
$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (\text{notation!})$$
- OLS Estimate, individual coefficient:

$$\hat{\beta}_j = \frac{\langle \mathbf{Y}, \mathbf{X}_{j\bullet} \rangle}{\|\mathbf{X}_{j\bullet}\|^2}$$

$\mathbf{X}_{j\bullet}$ = column $\mathbf{X}_j$ orthogonalized wrt all other $\mathbf{X}_k, k \neq j$

Statistical Significance in Linear Models

- Significance test of null hypothesis $H_0: \beta_j = 0$

Is $\hat{\beta}_j$ sufficiently large in magnitude to reject H_0 ?

$$t_j = \frac{(\hat{\beta}_j - 0)}{SE_j} = \frac{1}{\sigma} \langle \mathbf{Y}, \frac{\mathbf{X}_{j\bullet}}{\|\mathbf{X}_{j\bullet}\|} \rangle \stackrel{H_0}{=} \langle \frac{\epsilon}{\sigma}, \frac{\mathbf{X}_{j\bullet}}{\|\mathbf{X}_{j\bullet}\|} \rangle \sim \mathcal{N}_1(0, 1)$$

$\Rightarrow$ Reject H_0 if $|t_j| > K$, where $\mathbf{P}[|t_j| \leq K] \stackrel{H_0}{=} 1 - \alpha$ (e.g., 0.01)

- Confidence intervals with coverage $1 - \alpha$:

$$CI_j = [\hat{\beta}_j - K \cdot SE_j, \hat{\beta}_j + K \cdot SE_j]$$

This is a random interval that contains the true β_j with probability $1 - \alpha$.

- Typical: $K \approx 2$ for $\alpha \approx 0.05$

Submodels — Notation, Parameters, Assumptions

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

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

- The LS estimators in the submodel M are

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

- Coefficient estimates using orthogonalized columns in submodel M :

$$\hat{\beta}_{j \bullet M} = \frac{\langle \mathbf{Y}, \mathbf{X}_{j \bullet M} \rangle}{\|\mathbf{X}_{j \bullet M}\|^2}, \quad \mathbf{X}_{j \bullet M} \perp \mathbf{X}_k \forall k \neq j, \quad \mathbf{X}_{j \bullet M} \in \text{cspan}(\mathbf{X}_M)$$

- Associated t -statistics:

$$t_{j \bullet M} = \left\langle \frac{\mathbf{Y}}{\sigma}, \frac{\mathbf{X}_{j \bullet M}}{\|\mathbf{X}_{j \bullet M}\|} \right\rangle \stackrel{H_0}{=} \left\langle \frac{\epsilon}{\sigma}, \frac{\mathbf{X}_{j \bullet M}}{\|\mathbf{X}_{j \bullet M}\|} \right\rangle$$

Geometry of Adjustment Vectors

- The p vectors $\mathbf{X}_1, \dots, \mathbf{X}_p$ generate the “adjustment set”

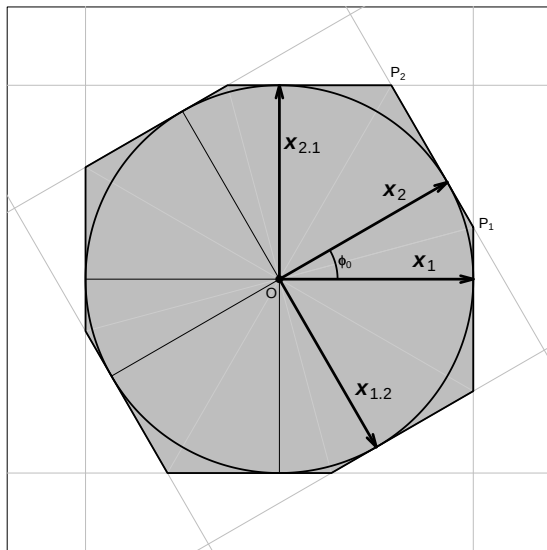
$$((\mathbf{X}_{j \bullet M} / \|\mathbf{X}_{j \bullet M}\|)_{j \in M})_M \quad (\text{Note: } \mathbf{X}_{j \bullet M} \in \text{cspan}(\mathbf{X}_M) \subset \text{cspan}(\mathbf{X}))$$

- Their number is $p2^{p-1}$. (If $\mathbf{X}_j \perp \mathbf{X}_k \forall j \neq k$, their number is p .)
- Gram-Schmidt series, example: $\mathbf{X}_1 \perp \mathbf{X}_{2 \bullet \{1,2\}} \perp \mathbf{X}_{3 \bullet \{1,2,3\}} \perp \dots$
- Orthogonalities in general: $\mathbf{X}_{j \bullet M} \perp \mathbf{X}_{j' \bullet M'}$ if $M \subset M', j \in M, j' \in M' \setminus M$
- Each $\mathbf{X}_{j \bullet M}$ is orthogonal to $(p-1)2^{p-2}$ adjustment vectors.
- **A Duality:**

Let the p vectors $\mathbf{X}_{1 \bullet F}, \dots, \mathbf{X}_{p \bullet F}$ be the adjustment vectors in the full model $F = \{1, 2, \dots, p\}$.

Fact: $\mathbf{X}_{1 \bullet F}, \dots, \mathbf{X}_{p \bullet F}$ generate the same adjustment set as $\mathbf{X}_1, \dots, \mathbf{X}_p$,
that is: $((\mathbf{X}_{j \bullet M})_{j \in M} / \|\mathbf{X}_{j \bullet M}\|)_M$.

Geometry of Adjustment — Normalized



Column space
of $\mathbf{X}$ for $p=2$
predictors,
partly collinear

Submodels versus the Full Model — Confusions

- Questions:

- ▶ How do submodel estimates $\hat{\beta}_M$ relate to full-model estimates $\hat{\beta}_{Full}$?
- ▶ How do submodel parameters β_M relate to full-model parameters β_{Full} ?

Is $\hat{\beta}_M$ a subset of $\hat{\beta}_{Full}$ and β_M a subset of β ?

- Answer: Unless $\mathbf{X}$ is an orthogonal design,

- ▶ $\hat{\beta}_M$ is **not** a subset of $\hat{\beta}_{Full}$ and
- ▶ β_M is **not** a subset of β_{Full} .

- Reason: Slopes, both estimates and parameters, depend on what the other predictors are — **in value and in meaning**.

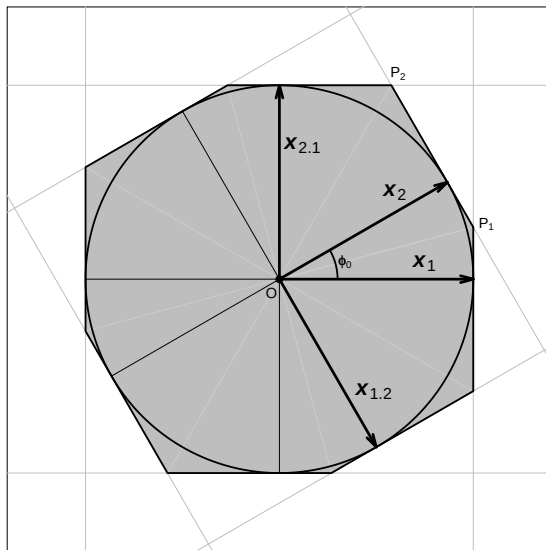
- Message: $\hat{\beta}_M$ does **not** estimate full-model parameters!

(Care is needed when the full model is thought to be causal:
Submodel estimates may suffer from “omitted variables bias.”)

'Slopes depend on ...' — An Illustration

- Survey of potential purchasers of a new high-tech gizmo:
 - ▶ Response: “LoP” = Likelihood of Purchase (self-reported on a Likert scale)
 - ▶ Predictor 1: Age
 - ▶ Predictor 2: Income
- We expect: Younger customers have higher LoP, that is, $\beta_{\text{Age}} < 0$.
- Analysis shows:
 - ▶ Regression of LoP on Age: Older customers have higher LoP, $\beta_{\text{Age}} > 0$.
 - ▶ Regression of LoP on Age **and** Income:
Adjusted for Income, younger customers have higher LoP, $\beta_{\text{Age} \bullet \text{Income}} < 0$.
- A case of Simpson's paradox: $\beta_{\text{Age}} > 0 > \beta_{\text{Age} \bullet \text{Income}}$.
 - ▶ The **marginal** and the Income-**adjusted** slope have very **different values** and **different meanings**.
 - ▶ Neither of the regressions is “wrong”; both can be useful.

Geometry of Adjustment — Normalized



Column space
of $\mathbf{X}$ for $p=2$
predictors,
partly collinear

Variable Selection

- What is a variable selection procedure?

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

- $\hat{\mathbf{M}}$ 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, $\mathbf{H}$, ...).

- Facing up to post-selection inference: Confusers!

- Target of Inference: the vector $\beta_{\hat{\mathbf{M}}(\mathbf{Y})}$, its components $\beta_{j \cdot \hat{\mathbf{M}}(\mathbf{Y})}$ for $j \in \hat{\mathbf{M}}(\mathbf{Y})$.
- The target of inference is **random**.
- The target of inference has a **random dimension**: $\beta_{\hat{\mathbf{M}}(\mathbf{Y})} \in \mathbb{R}^{|\hat{\mathbf{M}}(\mathbf{Y})|}$
- Conditional on $j \in \hat{\mathbf{M}}$, the target component $\beta_{j \cdot \hat{\mathbf{M}}(\mathbf{Y})}$ has **random meanings**.
- When $j \notin \hat{\mathbf{M}}$ both $\beta_{j \cdot \hat{\mathbf{M}}}$ and $\hat{\beta}_{j \cdot \hat{\mathbf{M}}}$ are **undefined**.
- Hence the coverage probability $\mathbf{P}[\beta_{j \cdot \hat{\mathbf{M}}} \in \text{CI}_{j \cdot \hat{\mathbf{M}}}(K)]$ is **undefined**.

Universal Post-Selection Inference

- Candidates for meaningful coverage probabilities:

- ▶ Marginal coverage: Berk et al. (2013)

$$\mathbf{P}[\beta_{j \cdot \hat{M}} \in \text{CI}_{j \cdot \hat{M}}(K) \quad \forall j \in \hat{M}]$$

- ▶ Conditional coverage: Lee et al. (2016) for $\hat{M} = \text{lasso}$

$$\mathbf{P}[\beta_{j \cdot M} \in \text{CI}_{j \cdot M} \mid \hat{M} = M] \quad \text{for } j \in M$$

- Overcoming a difficulty in marginal coverage:

- ▶ Problem: We don't know how to control marginal coverage for specific model selection procedures $\hat{M}$.
- ▶ Solution: Ask for more!

Universal post-selection coverage **for all selection procedures**

Reduction to Simultaneous Inference

Lemma

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

$$\max_{j \in \hat{M}} |t_{j \cdot \hat{M}}| \leq \max_{\mathbf{M}} \max_{j \in \mathbf{M}} |t_{j \cdot \mathbf{M}}| \quad \forall \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^N.$$

Theorem

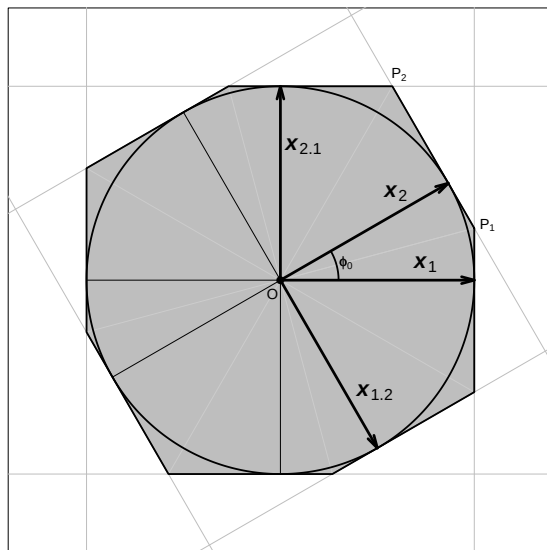
Let K be the $1 - \alpha$ quantile of the “max-max- $|t|$ ” statistic of the lemma:

$$\mathbf{P} \left[\max_{\mathbf{M}} \max_{j \in \mathbf{M}} |t_{j \cdot \mathbf{M}}| \leq K \right] \stackrel{(\geq)}{=} 1 - \alpha.$$

Then we have the following universal PoSI guarantee:

$$\mathbf{P} \left[\beta_{j \cdot \hat{M}} \in Cl_{j \cdot \hat{M}}(K) \quad \forall j \in \hat{M} \right] \geq 1 - \alpha \quad \forall \hat{M}.$$

PoSI Geometry — Simultaneity



PoSI polytope
= intersection
of all t -bands.

A Selection Method Requiring Full PoSI Protection

- Is there a model selection procedure that requires full PoSI protection?
- Consider $\hat{M}$ defined as follows:

$$\hat{M} := \operatorname{argmax}_M \max_{j \in M} |t_{j \cdot M}|$$

$\Rightarrow$ “Significance Hunting”, a special case of “p-hacking” (SNS 2011)

Computing PoSI

- The simultaneity challenge: there are $p2^{p-1}$ statistics $|t_{j \bullet M}|$.

p	1	2	3	4	5	6	7	8	9	10
$\# t_{j \bullet M} $	1	4	12	32	80	192	448	1,024	2,304	5,120
p	11	12	13	14	15	16	17	18	19	20
$\# t_{j \bullet M} $	11,264	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 of K_{PoSI} in R, brute force, for $p \lesssim 20$.
- Computations are specific to a design $\mathbf{X}$: $K_{\text{PoSI}} = K_{\text{PoSI}}(\mathbf{X}, \alpha, df)$
- Computations depend only on the inner product matrix $\mathbf{X}^T \mathbf{X}$.
 $\Rightarrow$ The limiting factor is p (N may only matter for $\hat{\sigma}^2$).
- One Monte Carlo computation is good for any α and any error df .
- R package **PoSI**

We don't use the special structure of the adjustment set!

Any Ideas?

Scheffé Protection Yields Valid PoSI

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

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

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

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

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

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

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

Orthogonal Designs have the Smallest PoSI Constants

- In orthogonal designs there is no adjustment:

$$\mathbf{X}_{j \cdot M}^{\text{orth}} = \mathbf{X}_j^{\text{orth}} \quad \forall M, j (\ni M)$$

- The PoSI statistic simplifies to $\max_{j=1 \dots p} |t_{j \cdot \{j\}}|$,
hence the PoSI guarantee reduces to

simultaneity for p orthogonal contrasts.

- The PoSI constant for orthogonal designs is uniformly smallest:

$$K_{\text{orth}}(p, \alpha, df) \leq K_{\text{PoSI}}(\mathbf{X}_{\dots \times p}, \alpha, df) \quad \forall p, \alpha, df, \mathbf{X}_{\dots \times p}$$

PoSI Asymptotics

- Natural asymptotics for the PoSI constant $K_{\text{PoSI}}(\mathbf{X}_{\dots \times p}, \alpha, df)$ are in terms of design sequences $p \mapsto \mathbf{X}_{\dots \times p}$ as $p \uparrow \infty$ and $df = \infty$, i.e., σ known.

- The Scheffé constant has the following rate in p :

$$K_{\text{Sch}}(p, \alpha) = \sqrt{\chi_{p; 1-\alpha}^2} \sim \sqrt{p}.$$

- ▶ This represents an upper bound on the PoSI rate.
- ▶ We know a sharper rate bound to be $0.866 \dots \sqrt{p}$.
- ▶ We know of design sequences that reach $0.78 \dots \sqrt{p}$.

- The lowest rate is achieved by orthogonal designs with a rate

$$K_{\text{orth}}(p, \alpha) \sim \sqrt{2 \log p}.$$

- Hence there is a wide range of rates for the PoSI constants:

$$\sqrt{2 \log p} \lesssim K_{\text{PoSI}}(\mathbf{X}_{\dots \times p}, \alpha) \lesssim \sqrt{p}$$

$$p = 100 : \quad \sqrt{2 \log p} \approx 3, \quad \sqrt{p} = 10$$

Worst Case PoSI Asymptotics — Some Details

- To achieve rates $\sim \sqrt{p}$, it is sufficient to consider “PoSI1”:
 - ▶ Protect $\mathbf{X}_p$, i.e., keep it in all models.
 - ▶ Search only models $M \subset \mathcal{M} = \{1, \dots, p-1\}$.
 - ▶ Need only consider $\max_{M \in \mathcal{M}} |t_{p \cdot M}|$,
i.e., all possible adjustments of $\mathbf{X}_p$.
- Example that achieves $K \sim 0.78\sqrt{p}$:

$$\begin{vmatrix} 1 & 0 & 0 & 0 & \dots & c \\ 0 & 1 & 0 & 0 & \dots & c \\ 0 & 0 & 1 & 0 & \dots & c \\ 0 & 0 & 0 & 1 & \dots & c \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \sqrt{1 - (p-1)c^2} \end{vmatrix}$$

$$c \uparrow \sqrt{1/(p-1)}$$

Ways to Limit the Size of the PoSI Problem

- The full universe of models for full PoSI: all non-singular submodels
 - ▶ $\mathcal{M}_{\text{all}} = \{M : M \subset \{1, 2, \dots, p\}, 0 < |M| \leq \min(n, p), \text{rank}(\mathbf{X}_M) = |M|\}.$
- Useful sub-universes:
 - (1) Protect one or more predictors, as in PoSI1: $\mathcal{M} = \{M : p \in M\}.$
 - (2) Sparsity, i.e., submodels of size m' or less: $\mathcal{M} = \{M : |M| \leq m'\}.$
 - (3) Richness, i.e., drop fewer than m' predictors from the full model:
 $\mathcal{M} = \{M : |M| \geq p - m'\}.$
 - (4) Nested sets of models, as in polynomial regression, AR models, ANOVA.
- Available:
 - (1) Protection: Orthogonalize wrt (adjust for) protected predictors.
 - (2,3) Limiting model sizes in the PoSI package: argument “modelSZ=1:5” (e.g.)

PoSI Significance: Strong Error Control

For each $j \in M$, consider the t -test statistic

$$t_{0,j \cdot \mathbf{M}} = \frac{\hat{\beta}_{j \cdot \mathbf{M}} - 0}{\hat{\sigma}_{\cdot} \cdot ((\mathbf{X}_{\mathbf{M}}^T \mathbf{X}_{\mathbf{M}})^{-1})_{jj}^{\frac{1}{2}}}.$$

Theorem

Let H_1 be the random set of true alternatives in $\hat{M}$,
and $\hat{H}_1$ the random set of rejections in $\hat{M}$:

$$\hat{H}_1 = \{(j, \hat{M}) : j \in \hat{M}, |t_{0,j,\hat{M}}| > K\} \quad \text{and} \quad H_1 = \{(j, \hat{M}) : j \in \hat{M}, \beta_{j,\hat{M}} \neq 0\}.$$

Then

$$\mathbb{P}(\hat{H}_1 \subset H_1) \geq 1 - \alpha.$$

If we repeat the sampling process many times, the probability that all PoSI rejections are correct is at least $1 - \alpha$, no matter how the model is selected.

Example: Length of Criminal Sentence (contd.)

- Reminder: t -statistics of selected covariates, in descending order:
 - ▶ $|t_{\text{alcohol}}| = 3.95$;
 - ▶ $|t_{\text{prior records}}| = 3.59$;
 - ▶ $|t_{\text{seriousness}}| = 3.57$;
 - ▶ $|t_{\text{drugs}}| = 3.31$;
 - ▶ $|t_{\text{employment}}| = 3.04$;
 - ▶ $|t_{\text{initial age}}| = 2.56$;
 - ▶ $|t_{\text{gender}}| = 2.33$.
- The PoSI constant is $K_{\text{PoSI}} \approx 3.1$, hence we would claim significance for the four variables on the left.
- For comparison, the Scheffé constant is $K_{\text{Sch}} \approx 4.5$, leaving us with no significant predictors at all.
- Similarly, Bonferroni with $\alpha/(p 2^{p-1})$ yields $K_{\text{Bonf}} \approx 4.7$.

Some Nice Features of PoSI

- PoSI solves a circularity problem:
 - ▶ You selected a model with all-subset search and C_p , for example.
 - ▶ You find some insignificant predictors in it.
 - ▶ You remove them – now what?

⇒ The PoSI multiplier still yields valid inference in the reduced model.
- Informal and post hoc selection is ok.
- One computation provides instant access to multipliers K of the standard error for all α .

Dissatisfaction with PoSI: Remaining Assumptions

- Homoskedastic Gaussian errors around a general mean vector
- The predictors are fixed, not random. If random, condition on them.

Why?

- In regression, are errors the only source of sampling variability?
- Is it permissible to condition on the predictors?

Evidence that the answers are “No”:

“The Unconditional Movie”

Execute in R the following line of code:

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

Following up on the consequences of these hints would be the subject of another talk.

Conclusions

- Valid marginal universal post-selection inference is possible.
- PoSI is not procedure-specific, hence is conservative. However:
 - ▶ PoSI is valid even for selection that is informal and post-hoc.
 - ▶ PoSI is necessary for selection based on “significance hunting”.
- Asymptotics in p suggests strong dependence of K_{PoSI} on design $\mathbf{X}$.
- **Unsolved:**
 - ▶ **Understanding the design geometry that drives $K_{\text{PoSI}}(\mathbf{X})$.**
 - ▶ **Computations of K_{PoSI} for large p .**
 - ▶ **Ideally: Find a simple geometric characteristic of $\mathbf{X}$ that drives the size of K_{PoSI} .**

Conclusions

- Valid marginal universal post-selection inference is possible.
- PoSI is not procedure-specific, hence is conservative. However:
 - ▶ PoSI is valid even for selection that is informal and post-hoc.
 - ▶ PoSI is necessary for selection based on “significance hunting”.
- Asymptotics in p suggests strong dependence of K_{PoSI} on design $\mathbf{X}$.
- **Unsolved:**
 - ▶ **Understanding the design geometry that drives $K_{\text{PoSI}}(\mathbf{X})$.**
 - ▶ **Computations of K_{PoSI} for large p .**
 - ▶ **Ideally: Find a simple geometric characteristic of $\mathbf{X}$ that drives the size of K_{PoSI} .**

THANK YOU!