Post-selection inference: some answers and some questions

Arun Kumar Kuchibhotla

14 September, 2021

Carnegie Mellon University

Table of contents

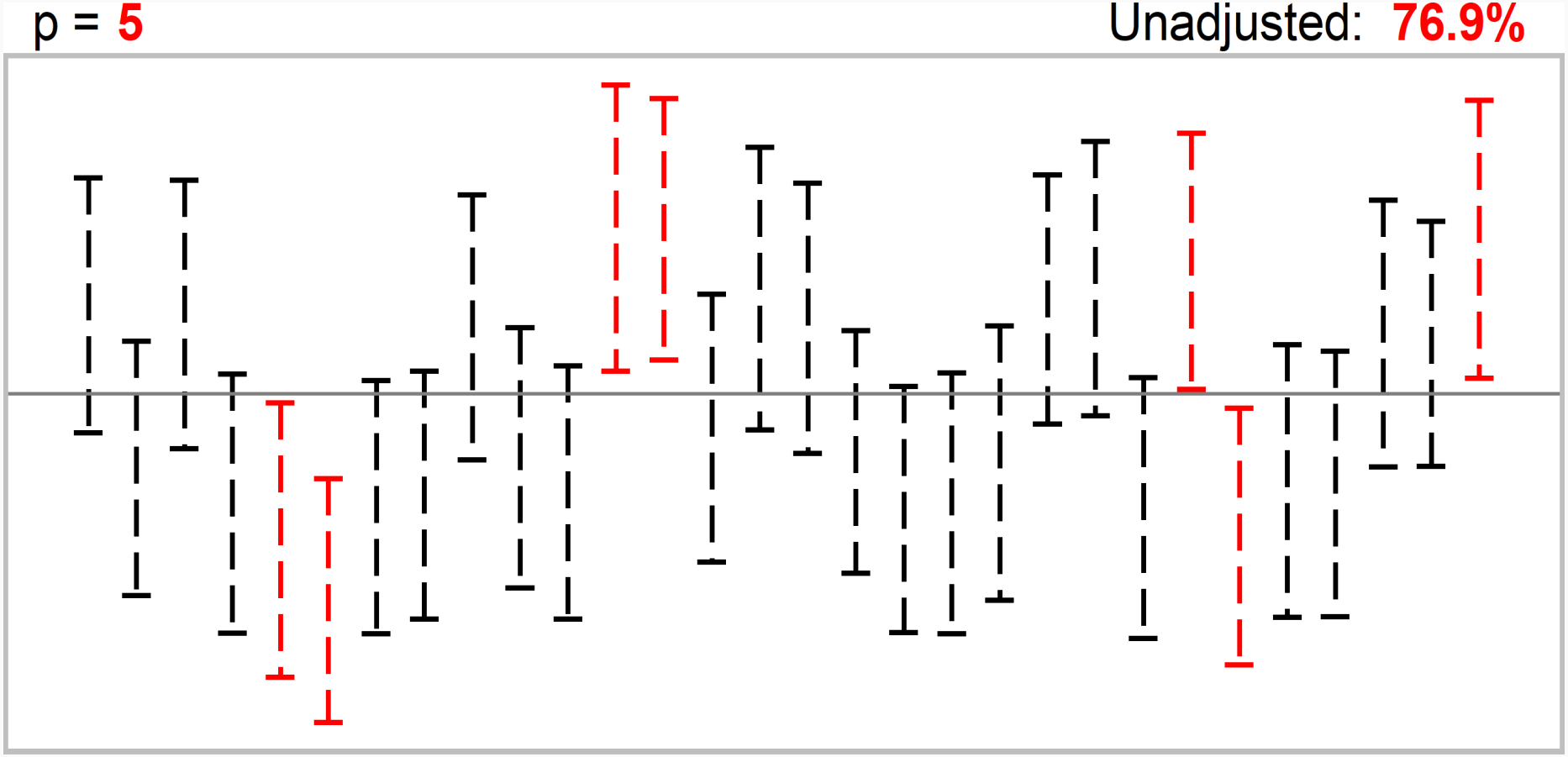
Invalidity of Classical Inference

Example: Invalidity of classical inference under selection

Generate 500 observations from $(X, Y) \sim N(0, I_{p+1})$. $(Y \perp X)$

Select one covariate $X_{\hat{j}}$ that is most correlated with Y.

Coverage of classical **95**% confidence interval

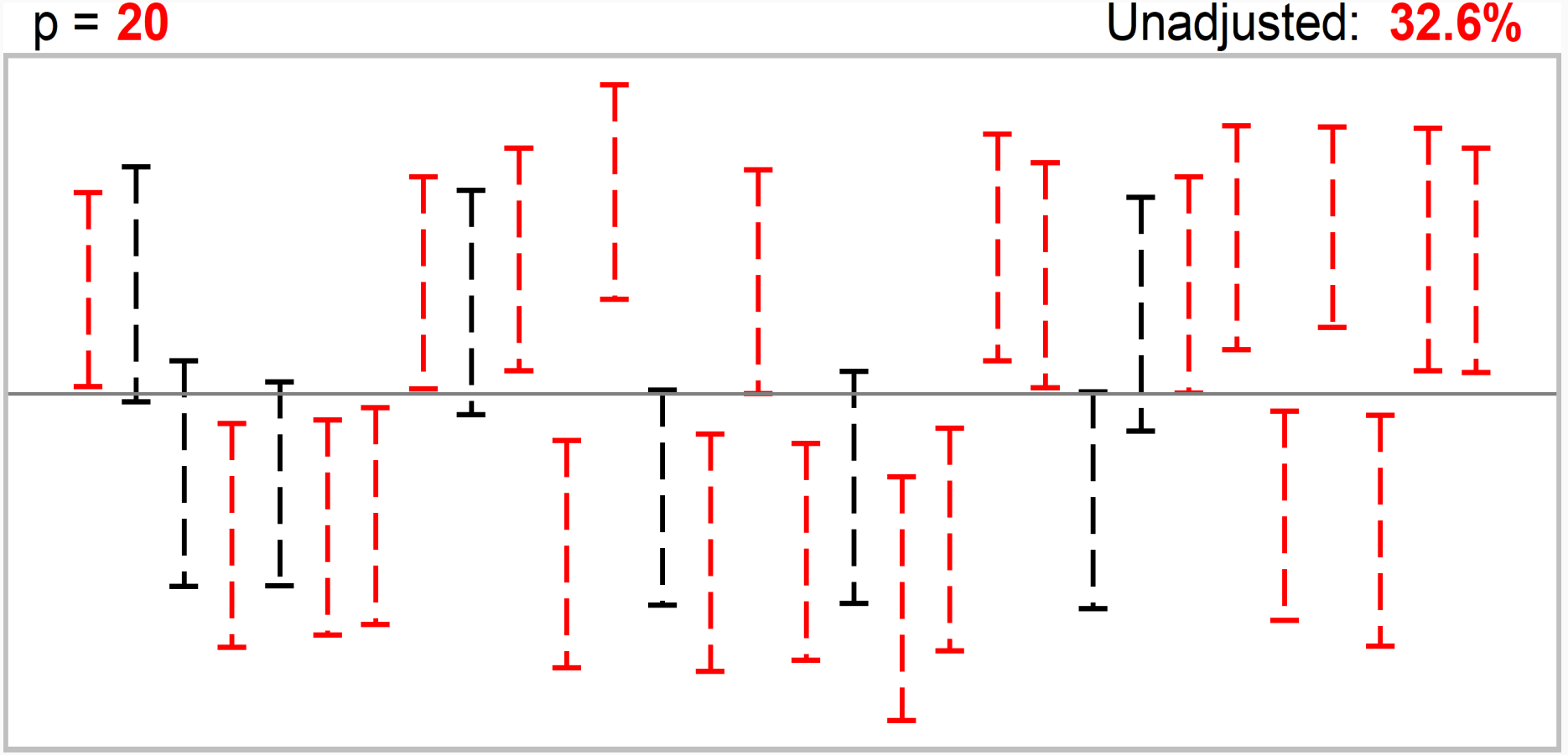


Example: Invalidity of classical inference under selection

Generate 500 observations from $(X, Y) \sim N(0, I_{p+1})$. $(Y \perp X)$

Select one covariate $X_{\hat{j}}$ that is most correlated with Y.

Coverage of classical **95**% confidence interval

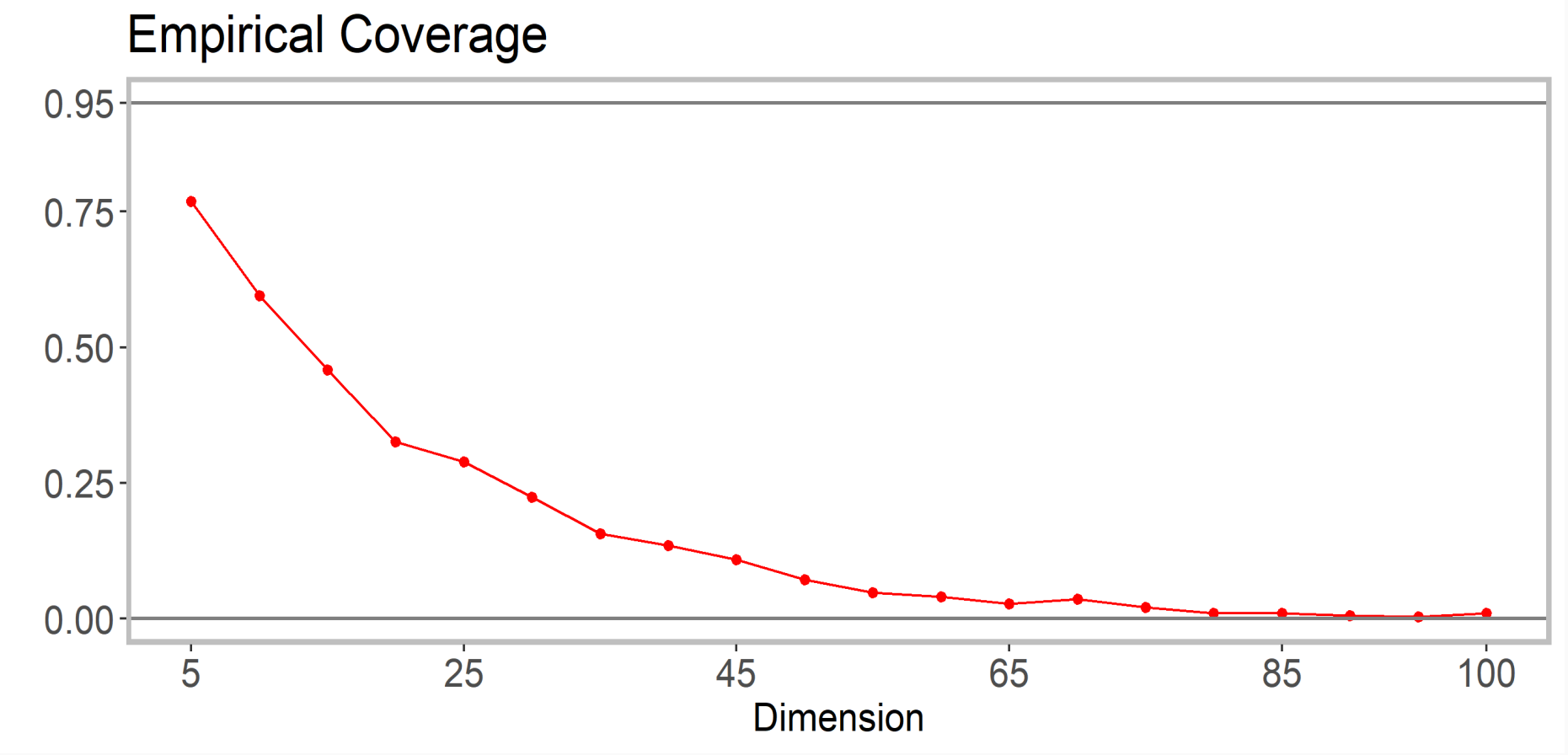


Example: Invalidity of classical inference under selection

Generate 500 observations from $(X, Y) \sim N(0, I_{p+1})$. $(Y \perp X)$

Select one covariate $X_{\hat{j}}$ that is most correlated with Y.

Coverage of classical **95%** confidence interval.



Summary

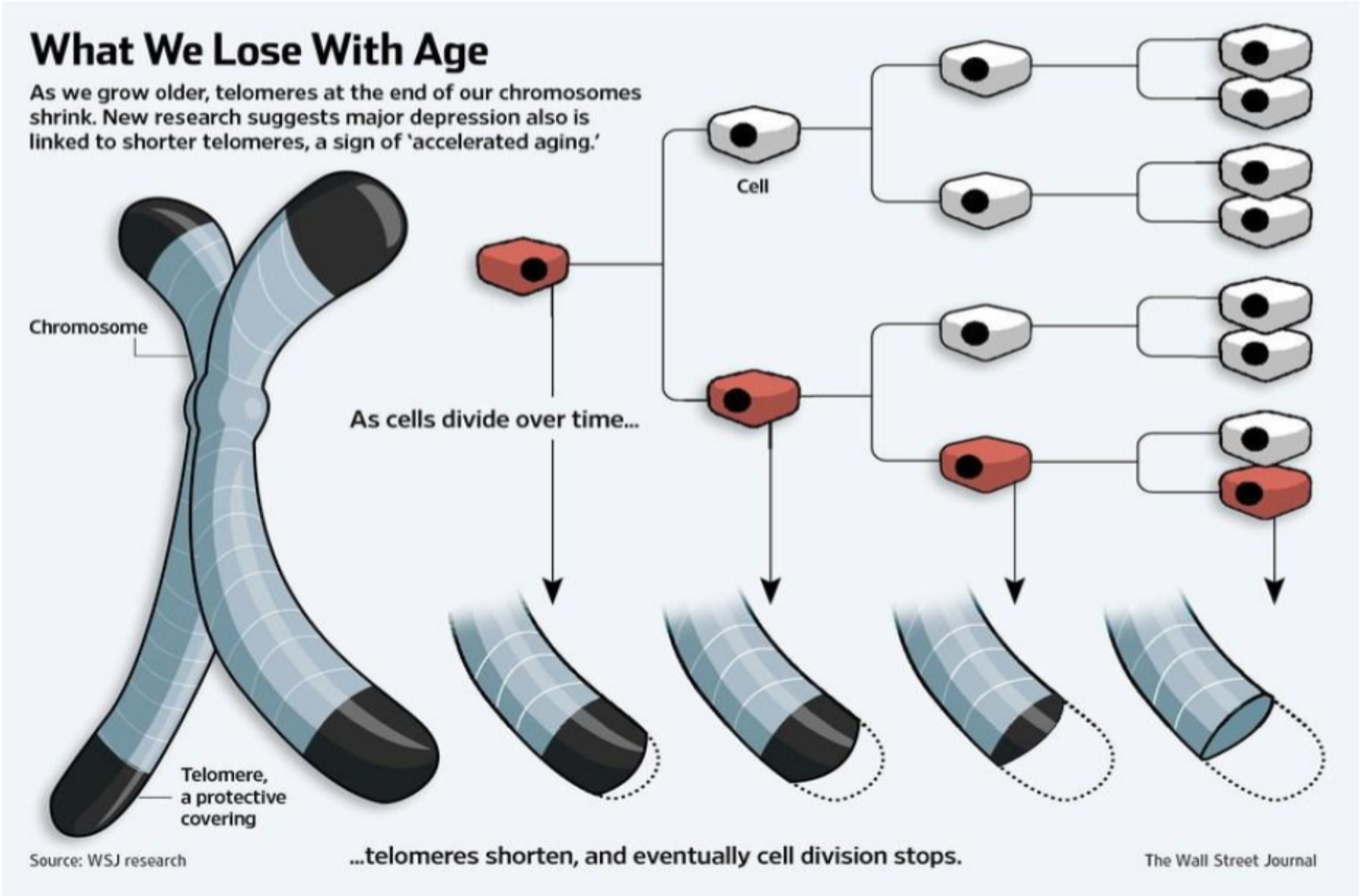
Unadjusted classical inference can be (very) misleading.

Duality of confidence intervals and testing implies that classical tests may not control Type I error.

It does not require a pathological selection to invalidate classical inference.

More concerningly, common practice of data exploration is very informal and imprecise.

Case Study 2: Covariate Selection



Case Study 2: Covariate Selection

nature.com Scientific Reports, 2019

WGS-based telomere length analysis in Dutch family trios implicates stronger maternal inheritance and a role for *RRM1* gene

"The MLR models were tested by sequential introduction of predictors and interaction terms.

• • •

ultimately, from the three best models with similar adjusted R squared values the simplest one was chosen."

Formulation of the Problem

There are p hypotheses to start with

$$H_{0,j} \ : \ \text{corr}(Y, X_j) \ = \ 0, \quad \text{for} \quad 1 \leq j \leq p.$$

Equivalently,

$$H_{0,j} \ : \ \beta_j \ = \ 0, \quad \text{for} \quad 1 \leq j \leq p,$$

where

$$(\alpha_j, \beta_j) \ := \ \underset{(\alpha,\beta)}{\text{argmin}} \ \mathbb{E}\left[(Y - \alpha - \beta X_j)^2\right].$$

There are p hypotheses to start with

$$H_{0,j} \ : \ \text{corr}(Y, X_j) \ = \ 0, \quad \text{for} \quad 1 \leq j \leq p.$$

Equivalently,

$$H_{0,j} \ : \ \beta_j \ = \ 0, \quad \text{for} \quad 1 \leq j \leq p,$$

where

$$(\alpha_j, \beta_j) \ := \ \underset{(\alpha,\beta)}{\text{argmin}} \ \mathbb{E}\left[(Y - \alpha - \beta X_j)^2\right].$$

Select a $\widehat{j} \in \{1, 2, \ldots, p\}$ based on the data.

There are p hypotheses to start with

$$H_{0,j} \ : \ \text{corr}(Y, X_j) \ = \ 0, \quad \text{for} \quad 1 \leq j \leq p.$$

Equivalently,

$$H_{0,j} \ : \ \beta_j \ = \ 0, \quad \text{for} \quad 1 \leq j \leq p,$$

where

$$(\alpha_j, \beta_j) \ := \ \underset{(\alpha,\beta)}{\text{argmin}} \ \mathbb{E}\left[(Y - \alpha - \beta X_j)^2\right].$$

Select a $\widehat{j} \in \{1, 2, \ldots, p\}$ based on the data.

Test the hypothesis $H_{0,\widehat{j}} : \beta_{\widehat{j}} = 0$.

Classical (invalid) test:

$$\text{Reject } H_{0,\widehat{j}} \text{ if } \quad |t_{\widehat{j}}| \ := \ \frac{n^{1/2}|\widehat{\beta}_{\widehat{j}}|}{\widehat{\sigma}_{\widehat{j}}} \ \leq \ 1.96.$$

The General PoSI Problem

For each model $\mathrm{M} \subseteq \{1, 2, \ldots, p\}$, define the OLS target as

$$\beta_{\mathrm{M}} := \underset{\theta \in \mathbb{R}^{|\mathrm{M}|}}{\operatorname{argmin}} \, \mathbb{E}\left[(Y - X_{\mathrm{M}}^{\top}\theta)^2\right].$$

Construct a confidence region $\widehat{\mathsf{CI}}_{\widehat{j}\cdot\widehat{\mathrm{M}}}$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\beta_{\widehat{j}\cdot\widehat{\mathrm{M}}} \in \widehat{\mathsf{CI}}_{\widehat{j}\cdot\widehat{\mathrm{M}}}\right) \geq 1 - \alpha,$$

for any model $\widehat{\mathrm{M}}$ (of size at most k) and $\widehat{j} \in \widehat{\mathrm{M}}$, irrespective of how it is chosen.

A Guiding Principle

Simultaneous Inference $\Rightarrow$ **Post-selection Inference**

$$\mathbb{P}\left(\bigcap_{\substack{|\mathrm{M}|\le k\\ j\in\mathrm{M}}}\left\{\beta_{j\cdot\mathrm{M}}\in\widehat{\mathsf{CI}}_{j\cdot\mathrm{M}}\right\}\right)\quad\le\quad\inf_{\substack{\widehat{j}\in\widehat{\mathrm{M}},\\ |\widehat{M}|\le k}}\mathbb{P}\left(\beta_{\widehat{j}\cdot\widehat{\mathrm{M}}}\in\widehat{\mathsf{CI}}_{\widehat{j}\cdot\widehat{\mathrm{M}}}\right).$$

Theorem:

Simultaneous inference is necessary for valid PoSI.

Three Solutions

A (Very) Simple Solution

Apply Bonferroni procedure.

$$\mathbb{P}\left(\bigcap_{\substack{|\mathrm{M}|\leq k\\ j\in\mathrm{M}}}\left\{\beta_{j\cdot\mathrm{M}} \in \widehat{\mathsf{CI}}_{j\cdot\mathrm{M}}\right\}\right) \geq 1-\sum_{\substack{|\mathrm{M}|\leq k,\\ j\in\mathrm{M}}}\mathbb{P}\left(\beta_{j\cdot\mathrm{M}} \in \widehat{\mathsf{CI}}_{j\cdot\mathrm{M}}\right).$$

How many elements in the sum?

$$\sum_{\substack{|\mathrm{M}|\leq k,\\ j\in\mathrm{M}}} 1 = \sum_{s=1}^{k} s\binom{p}{s} \asymp \left(\frac{ep}{k}\right)^k.$$

Construct $1-\dfrac{\alpha}{(ep/k)^k}$ confidence intervals for individual coefficients.

Can be very conservative.

Second Simple Solution

For simultaneous inference, inflate the interval to

$$\widehat{\mathsf{CI}}_{j\cdot\mathrm{M}} := \left\{\theta \in \mathbb{R} : \left|\frac{n^{1/2}(\widehat{\beta}_{j\cdot\mathrm{M}} - \theta)}{\widehat{\sigma}_{j\cdot\mathrm{M}}}\right| \leq K_\alpha\right\},$$

with K_α, the $(1-\alpha)$ quantile of

$$\max_{|\mathrm{M}|\leq k,\, j\in\mathrm{M}} \left|\frac{n^{1/2}(\widehat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}}\right|.$$

Accounts for dependence.

Disadvantage of these Solutions

Bonferroni Solution:

$$\widehat{\mathsf{CI}}_{j\cdot\mathrm{M}}^{\texttt{Bonf}} := \left\{\theta \in \mathbb{R} : \left|\frac{n^{1/2}(\widehat{\beta}_{j\cdot\mathrm{M}} - \theta)}{\widehat{\sigma}_{j\cdot\mathrm{M}}}\right| \leq z_{\alpha/(2(ep/k)^k)}\right\}.$$

PoSI Solution:

$$\widehat{\mathsf{CI}}_{j\cdot\mathrm{M}}^{\texttt{PoSI}} := \left\{\theta \in \mathbb{R} : \left|\frac{n^{1/2}(\widehat{\beta}_{j\cdot\mathrm{M}} - \theta)}{\widehat{\sigma}_{j\cdot\mathrm{M}}}\right| \leq K_\alpha\right\}.$$

K_α usually grows with largest model size k.

Say, $k = 20$, then

width of intervals for model of size 2

$$\approx$$

width of intervals for model of size 20.

The Third Solution

Define

$$\text{(OLS Estimator)} \qquad \widehat{\beta}_{\mathrm{M}} := \underset{\theta \in \mathbb{R}^{|M|}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_{i,\mathrm{M}}^{\top} \theta)^2,$$

$$\text{(OLS Target)} \qquad \beta_{\mathrm{M}} := \underset{\theta \in \mathbb{R}^{|M|}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(Y_i - X_{i,\mathrm{M}}^{\top} \theta)^2\right].$$

For any $\mathrm{M} \subseteq \{1, 2, \ldots, p\}$, consider the confidence region

$$\widehat{\mathrm{CI}}_{\mathrm{M}}^{\texttt{UPoSI*}} := \left\{ \theta \in \mathbb{R}^{|\mathrm{M}|} : \|\widehat{\Sigma}_{\mathrm{M}}(\widehat{\beta}_{\mathrm{M}} - \theta)\|_{\infty} \leq C_{xy}(\alpha) + C_{xx}(\alpha)\|\theta\|_1 \right\}.$$

Then for any model $\widehat{\mathrm{M}}$ chosen based on the data,

$$\boxed{\mathbb{P}\left(\beta_{\widehat{\mathrm{M}}} \in \widehat{\mathrm{CI}}_{\widehat{\mathrm{M}}}^{\texttt{UPoSI*}}\right) \geq 1 - \alpha,}$$

if $C_{xy}(\alpha)$ and $C_{xx}(\alpha)$ denote the $(1-\alpha)$ joint quantiles of

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \{X_i Y_i - \mathbb{E}[X_i Y_i]\} \right\|_{\infty} \quad \text{and} \quad \left\| \frac{1}{n} \sum_{i=1}^{n} \{X_i X_i^{\top} - \mathbb{E}[X_i X_i^{\top}]\} \right\|_{\infty}.$$

Comparison of Volumes

Reference	**Leb**$(\widehat{\mathrm{CI}}_{\widehat{\mathbb{M}}})$	Design
Kuchibhotla et al. (2021, AoS)	$(\log p/n)^{\lvert\widehat{\mathbb{M}}\rvert/2}$	fixed
	$(\lvert\widehat{\mathbb{M}}\rvert \log p/n)^{\lvert\widehat{\mathbb{M}}\rvert/2}$	random
Berk et al. (2013) Bachoc et al. (2019) Kuchibhotla et al. (2021, Econ. Theory)	$(k \log(ep/k)/n)^{\lvert\widehat{\mathbb{M}}\rvert/2}$	fixed/random

Table 1: Volumes of Different PoSI Regions.

Some Questions

Question 1: Optimality of volume

- We have shown that simultaneous inference is necessary and sufficient for post-selection inference in the problem formulation here.
- If arbitrary model selection procedures are allowed and the analyst chooses the covariate subset $\widehat{\mathrm{M}}$, then what is the best possible confidence set for $\beta_{\widehat{\mathrm{M}}}$ in terms of volume?
- The current best available is $(|\widehat{\mathrm{M}}| \log(ep/|\widehat{\mathrm{M}}|)/n)^{|\widehat{\mathrm{M}}|/2}$. This follows from the third solution above and also the Hierarchical PoSI proposed in Kuchibhotla (2020, PhD Thesis).
- From the literature of sparse high-dimensional linear regression, this volume seems to be the best possible.

What is the smallest volume confidence set for $\beta_{\widehat{\mathrm{M}}}$ with an (asymptotic) coverage validity for arbitrary selection of $\widehat{\mathrm{M}}$?

Question 1: Optimality of volume (Contd.)

- A simpler problem to consider is the normal means setting.
- Suppose $X \sim N(\mu, \Sigma)$ in $\mathbb{R}^d$ with known Σ. Let $\mu = (\mu_1, \ldots, \mu_d)^\top$.
- Consider the problem of constructing a confidence interval for $\mu_{\widehat{j}}$ for a data-based selection $\widehat{j} \in \{1, 2, \ldots, d\}$.
- For any fixed $1 \leq j \leq d$, a valid confidence interval for μ_j has width that scales as $\sqrt{\mathrm{Var}(X_j)}$.
- For a data-based selection $\widehat{j}$, one can construct confidence intervals of width of order $\sqrt{\log(\widehat{j})}\, \max_j \sqrt{\mathrm{Var}(X_j)}|_{j=\widehat{j}}$.
- It is currently unknown what the smallest volume confidence set for this simpler setting.

Question 2: Dynamic PoSI

- The post-selection inference problem posed is written with respect to a universe.
- For example, in linear regression with covariate selection, we have the universe of selection to be set of all subsets of covariates.
- In practice, this is not how data-drive selection is done. The second model selection method is informed by the output of the first model selection method, i.e., dynamic selection.
- For example, one might decide to (univariate) marginal screening first, then depending how many p-values are small might decide where to threshold the p-values for selecting covariates.
- Without any control on either the dynamics or the output of selection, PoSI is *impossible*.
- Dynamic PoSI via randomized output is acheived by adaptive data analysis and stable algorithms. But these methods require specific model selection methods and cannot handle visual selection.

Question 2: Dynamic PoSI (Contd.)

- Is there a generic way to modify the output of an arbitrary model selection strategy to solve dynamic PoSI?
- This includes those selections based on residual plots, QQ plots, and so on.
- At present, a very crude solution is available where by the underlying data is randomized to solve this. But this results in a huge lose of information.
- Furthermore, the methods that are more precise randomize functions of data with the amount of randomization depending on underlying distribution parameters that are seldom available.
- For example, in differential privacy, Laplace mechanism noise depends on the sensitivity of the function of data which is often unavailable.

Question 3: Computable PoSI

- The problem of post-selection inference as posed with respect to a universe of selection has several solutions now.
- Except in the linear regression case, there is no computable solution, i.e., some inference procedure that is not NP-hard.
- This is somewhat similar to the ℓ_0 (sparsity) penalized linear regression.
- The statistics used to compute simultaneous confidence intervals for PoSI at present involve maximum over the universe.
- In the context of covaraite selection, this is the maximum over all (sparse) subsets of covariates, which is NP-hard to compute.

Is there a computable solution to PoSI for general linear models?

Question 3: Computable PoSI (Contd.)

- Currently, there is work in progress that provides a randomized algorithm to compute the approximate maximum over the universe of selection.
- These approximations yield valid, albeit conservative confidence intervals for PoSI.
- It is not obvious what would be the best approximation to the maximum for best performance in practice.
- It is also not clear if there is a trade-off between conservativeness and computation.
- Computation is the main hurdle for the practical use of the existing PoSI solutions at present.

Conclusions

Post-selection inference is an important topic of research given the current reproducibility crisis.

There exist several solutions in the current formulation of the PoSI problem.

Their practical use is riddled with several obstacles and resolving them would lead to (more) trustworthy practical data analysis.

Reference: Kuchibhotla et al. (2021) Valid Post-selection Inference in Model-free Linear Regression, Annals of Statistics.

Conclusions

Post-selection inference is an important topic of research given the current reproducibility crisis.

There exist several solutions in the current formulation of the PoSI problem.

Their practical use is riddled with several obstacles and resolving them would lead to (more) trustworthy practical data analysis.

Reference: Kuchibhotla et al. (2021) Valid Post-selection Inference in Model-free Linear Regression, Annals of Statistics.

Thank you!