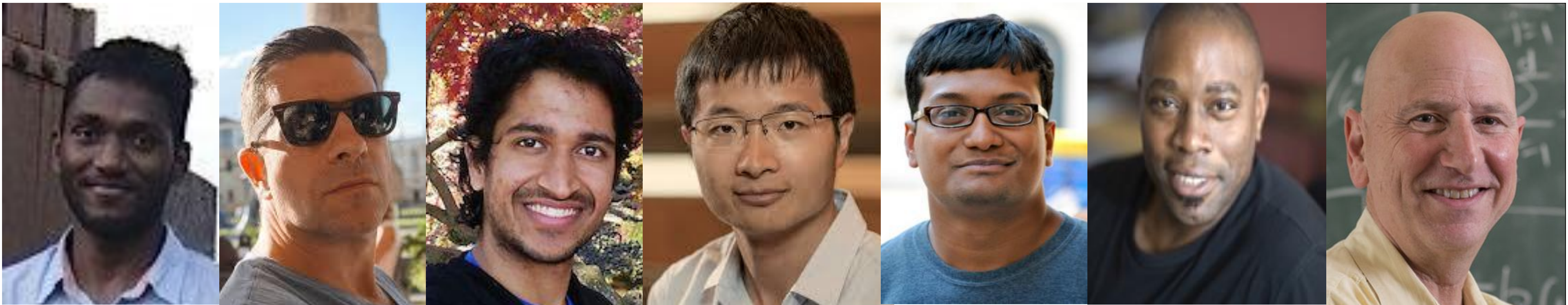
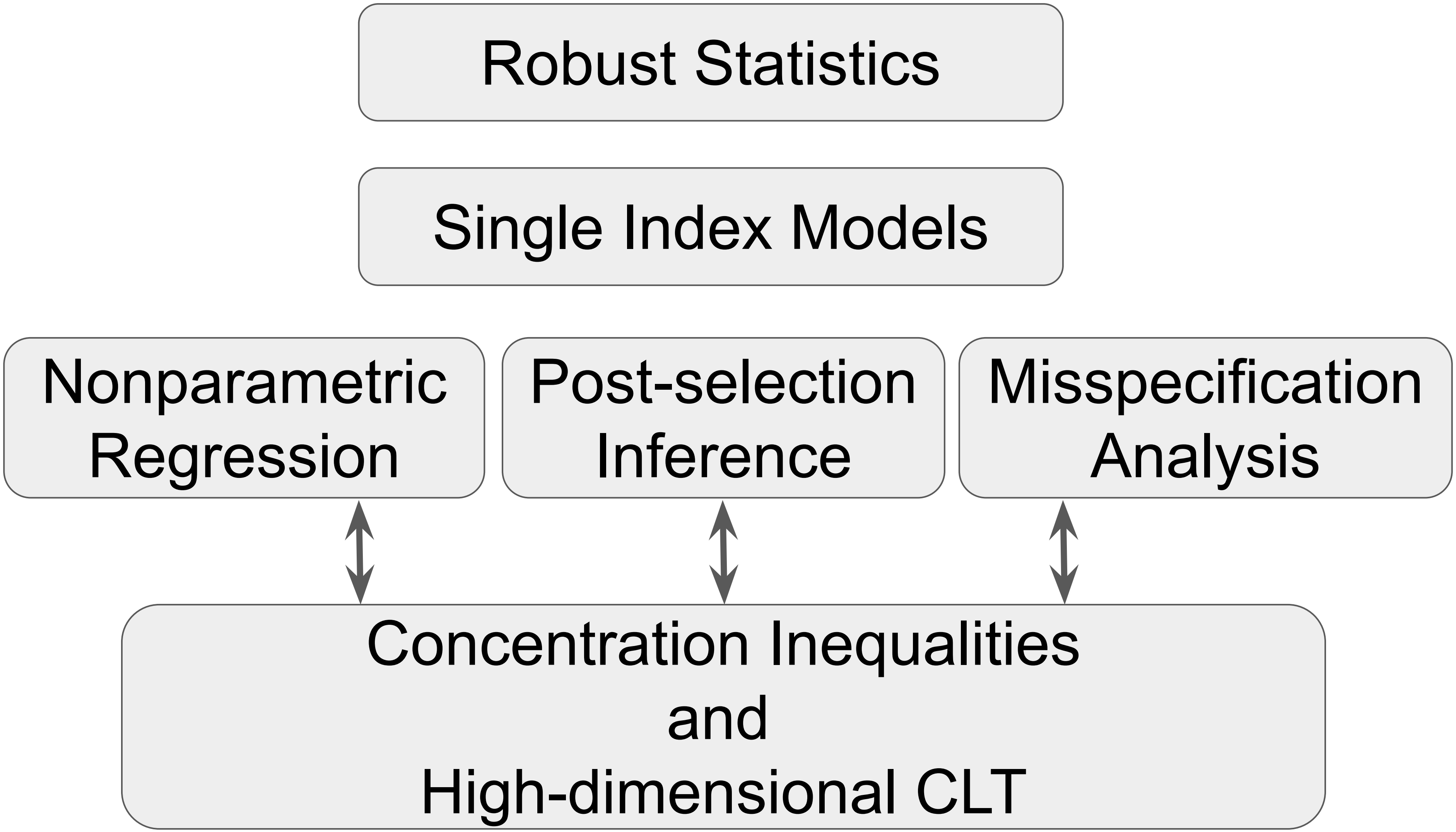
Arun Kumar Kuchibhotla

University of Pennsylvania

https://arun-kuchibhotla.github.io/



Overview of my research interests



Valid Post-selection Inference

Why and How

Arun Kumar Kuchibhotla
University of Pennsylvania

Larry Brown

Andreas Buja

Junhui Cai

Linda Zhao

Ed George

Roadmap

Data Snooping: Effects and Examples

Formulation of the Problem

Solution for Covariate Selection

Example and Conclusions

Roadmap

Data Snooping: Effects and Examples

- Effects illustrated with stepwise selection.
- Data snooping in textbooks and practice.

Formulation of the Problem

- The Problem & literature review for covariate selection.

Solution for Covariate Selection

- Key contributions
- Simulations & main components of the theory.

Example and Conclusions

- Real data example, Extensions & Summary

Roadmap

Data Snooping: Effects and Examples

- Effects illustrated with stepwise selection.
- Data snooping in textbooks and practice.

Formulation of the Problem

Solution for Covariate Selection

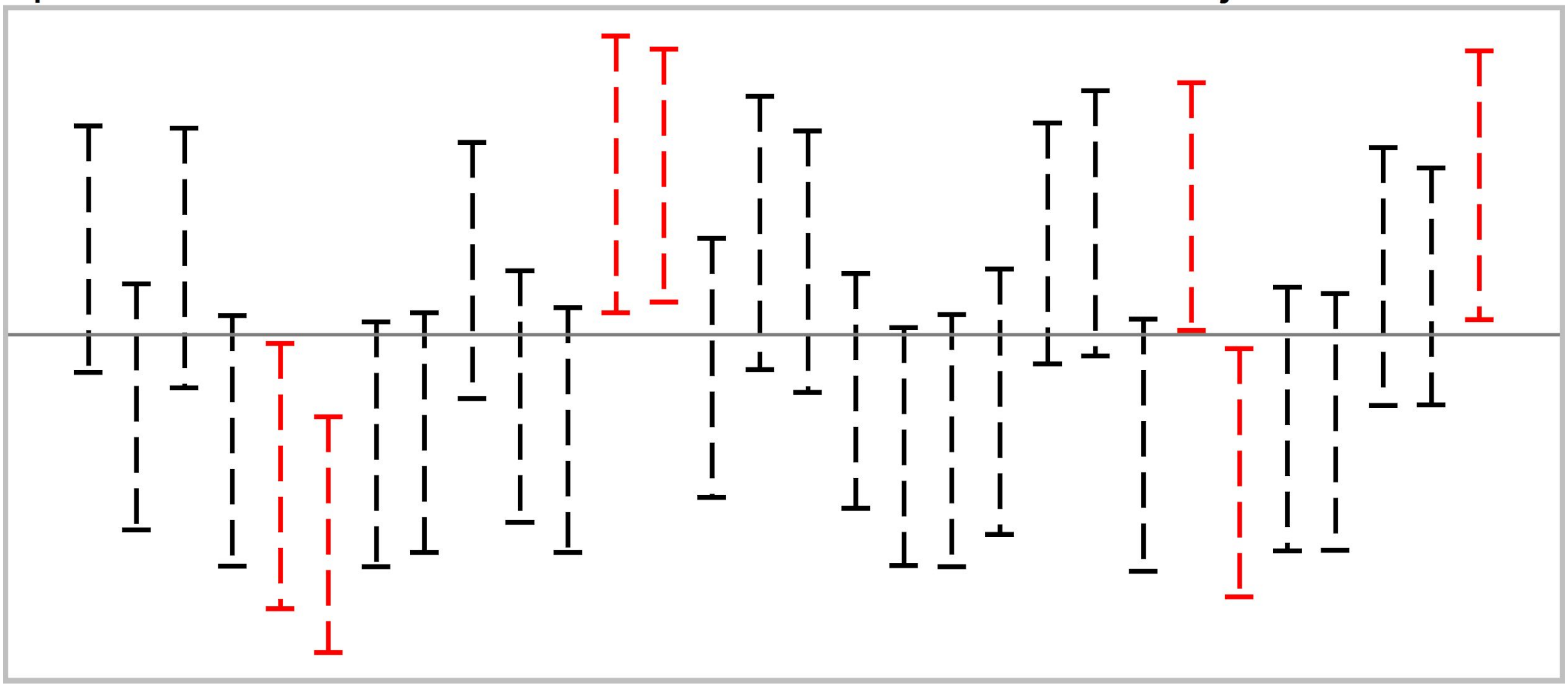
Example and Conclusions

Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1})$ $\Rightarrow$ 500 observations

$X_{\hat{j}}$ most correlated with Y $\Rightarrow$ 95% CI for slope

p = **5** Unadjusted: **76.9%**

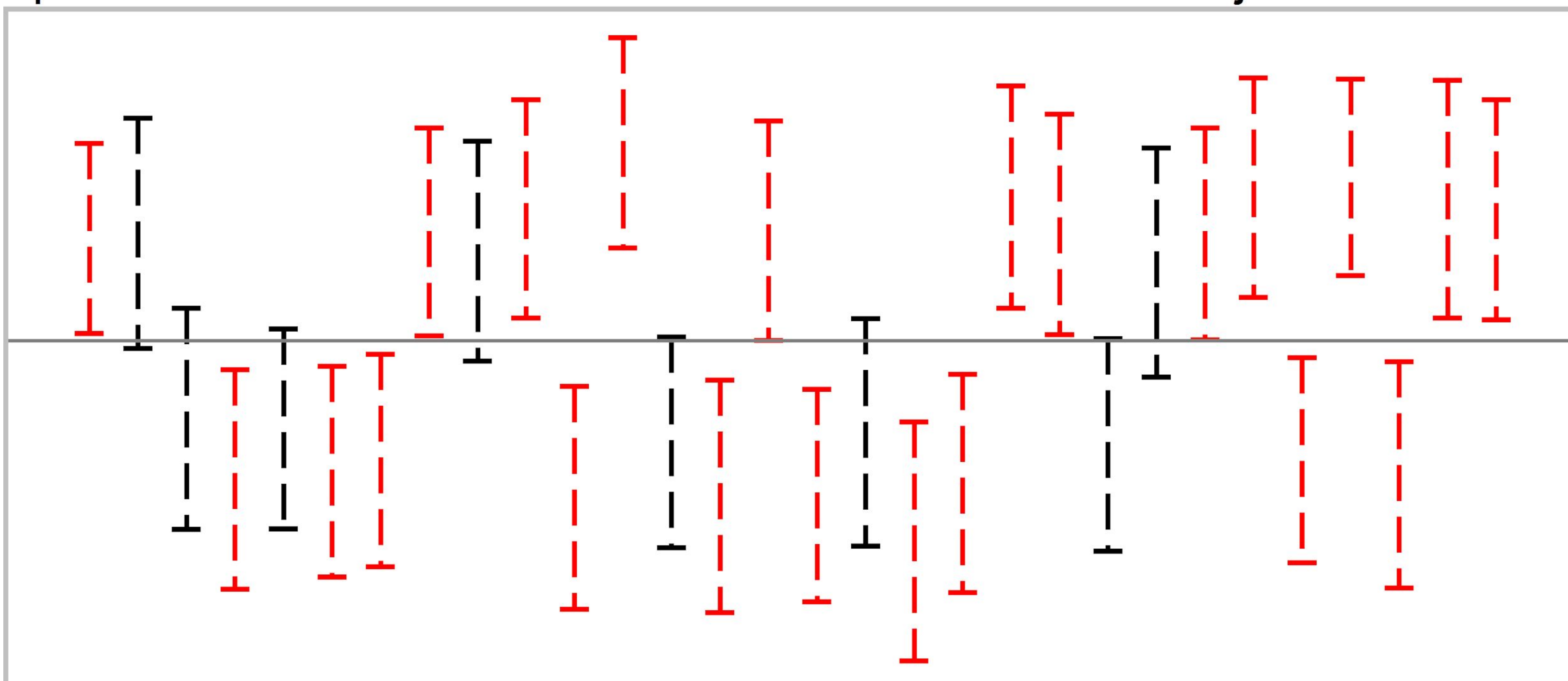


Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1})$ $\Rightarrow$ 500 observations

$X_{\hat{j}}$ most correlated with Y $\Rightarrow$ 95% CI for slope

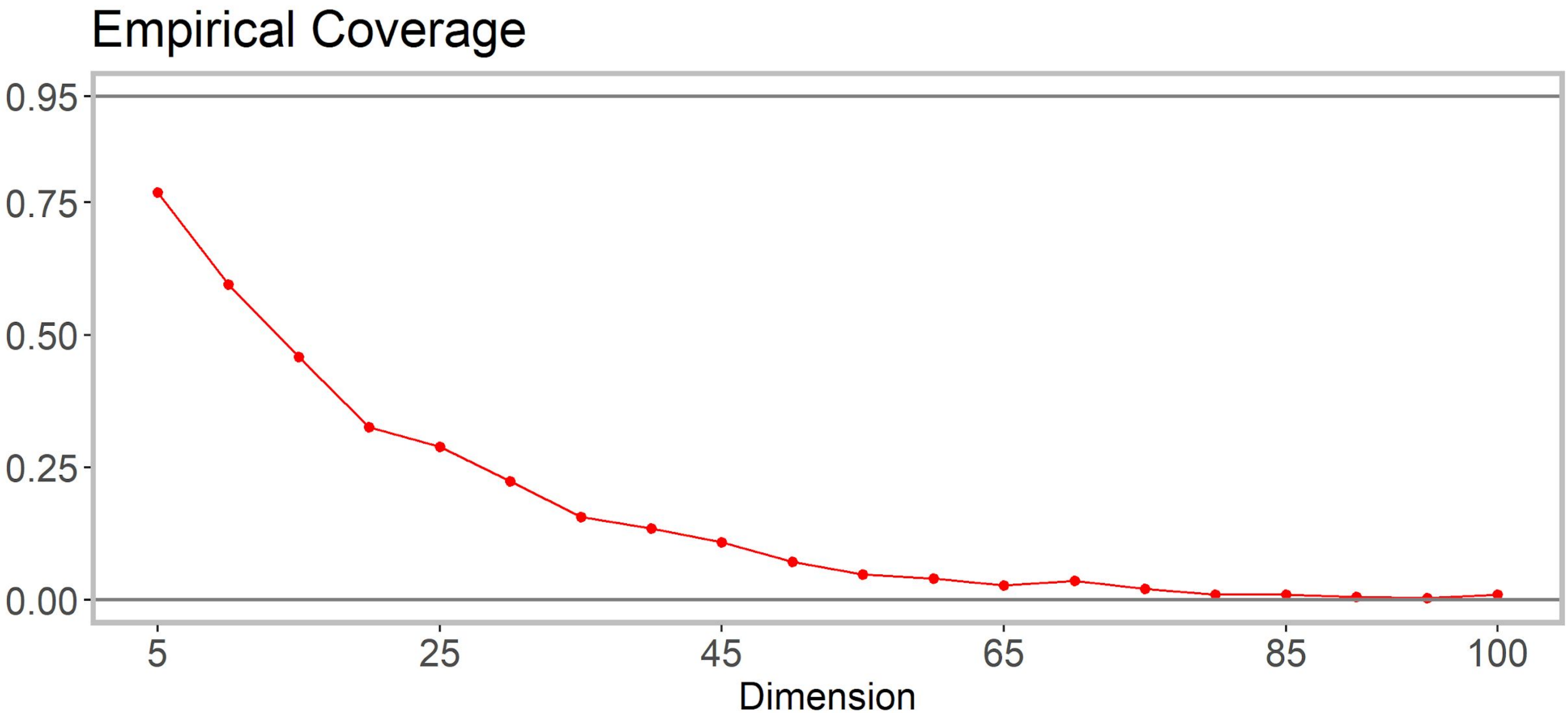
p = **20** Unadjusted: **32.6%**



Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1})$ $\Rightarrow$ 500 observations

$X_{\hat{j}}$ most correlated with Y $\Rightarrow$ 95% CI for slope



Some Notes

- Unadjusted inference after data snooping can be (very) misleading.
- Data snooping contributes to the

  Replicability Crisis

  - Inability to replicate conclusions in future studies.
  - 95% CIs should imply correct conclusions in 95% of studies.
- More concerningly, common practice of data snooping is more informal and imprecise than the example shown.

Case Study 1: Covariate Selection

British Medical Journal
2005

Postdischarge mortality in children with acute infectious diseases: derivation of postdischarge mortality prediction models

variate imputation using chained equations.[12] Following univariate analysis, candidate models were generated using a stepwise selection procedure minimising Akaike's Information Criterion (AIC). This method is

Case Study 1: Covariate Selection

British Medical Journal 2005

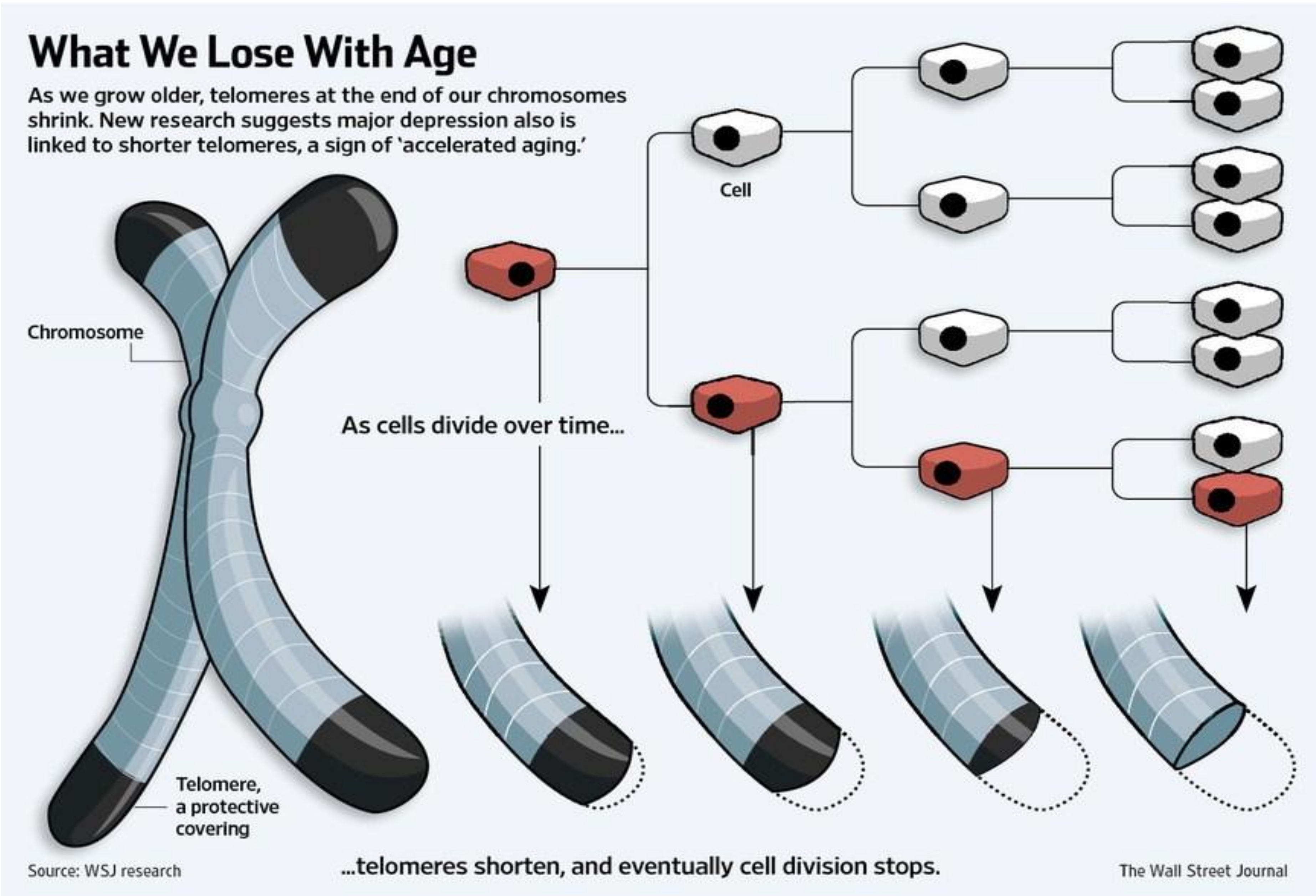
Postdischarge mortality in children with acute infectious diseases: derivation of postdischarge mortality prediction models

variate imputation using chained equations.[12] Following univariate analysis, candidate models were generated using a stepwise selection procedure minimising Akaike's Information Criterion (AIC). This method is

• • •

final selection of a model was judged on model parsimony (the simpler the better), availability of the predictors (with respect to minimal resources and cost), and the attained sensitivity (with at least 50% specificity). All

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."

Case Study 3: Transformations

Harrison and Rubinfeld (1978)★ write

to determine the best fitting functional form. Comparing models with either median value of owner-occupied homes (MV) or Log(MV) as the dependent variable, we found that the semilog version provided a slightly better fit. Using Log(MV) as the dependent variable, we concentrated on estimating a nonlinear term in NOX; i.e., we included NOX^p in the equation, where p is an unknown

The statistical fit in the equation was best when p was set equal to 2.0,

★H & R (1978) Hedonic housing prices and the demand for clean air.

Roadmap

Data Snooping: Effects and Examples

- Effects illustrated with stepwise selection.
- Data snooping in textbooks and practice.

Formulation of the Problem

- The Problem & literature review for covariate selection.

Solution for Covariate Selection

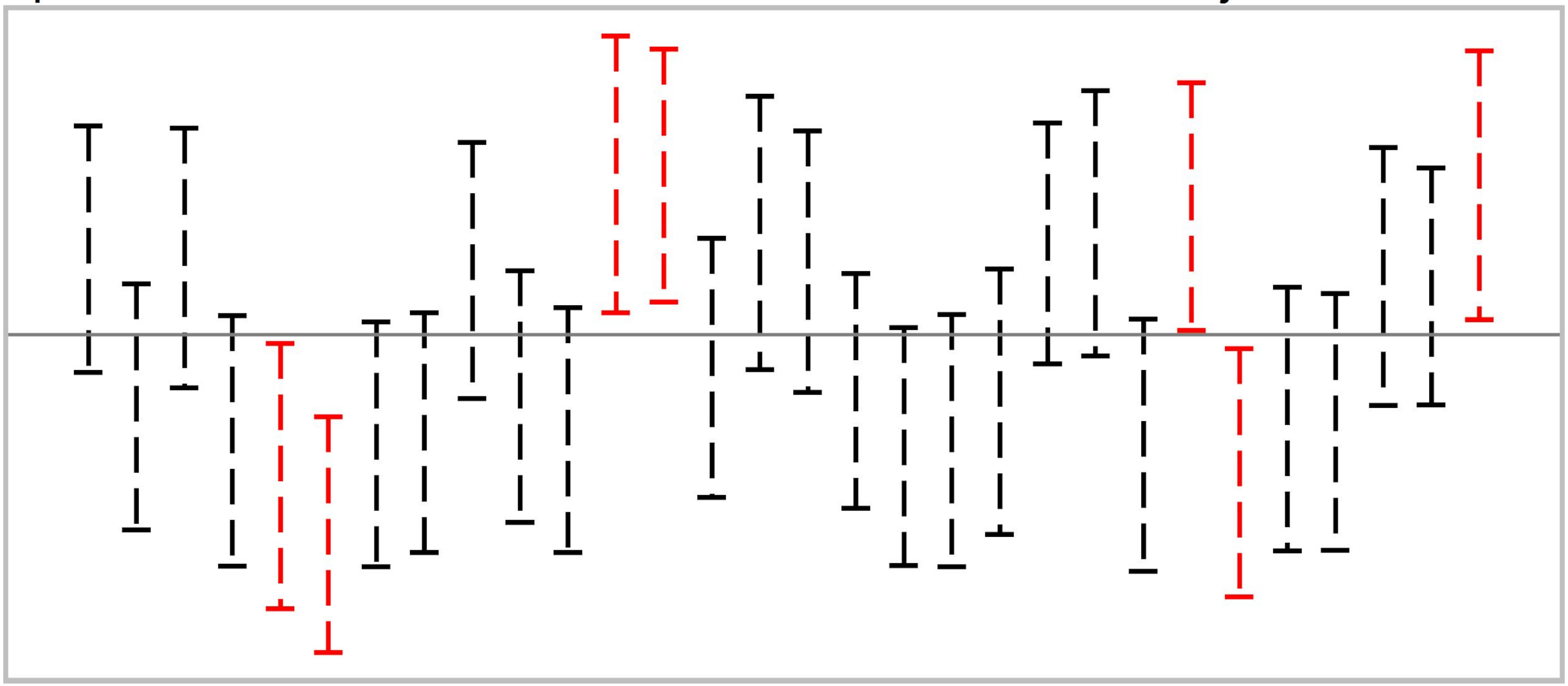
Example and Conclusions

Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1})$ $\Rightarrow$ 500 observations

$X_{\hat{j}}$ most correlated with Y $\Rightarrow$ 95% CI for slope

p = **5** Unadjusted: **76.9%**



Post-selection Inference: Problem 1

There are p covariates and for each $1 \leq j \leq p$

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

Want a valid CI for the parameter/target $\beta_{\widehat{j}}$:

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

irrespective of how $\widehat{j}$ is chosen based on data.

Post-selection Inference: Problem 2

For each $\mathrm{M} \subseteq \{1, 2, \dots, p\}$,

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

Want a valid CI for $\beta_{\hat{j}\cdot\widehat{\mathrm{M}}}$, a coordinate of $\beta_{\widehat{\mathrm{M}}}$:

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

irrespective of how $\widehat{\mathrm{M}}$ with size $\leq k$ and $\hat{j} \in \widehat{\mathrm{M}}$ are chosen based on data.

Solution 0: Sample Splitting

$\mathcal{D}_1 := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ | $\{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\} =: \mathcal{D}_2$

Variable selection, Transformations etc. | Finally, use once for inference.

✔ Allows arbitrary exploration in $\mathcal{D}_1$.

✘ Cannot revise the model after using $\mathcal{D}_2$.

✘ Invalidity when using multiple splits.

✘ Applicable only for independent data.

Rinaldo et al. (2019) Bootstrapping and sample splitting for high-dimensional, assumption-lean inference, *Annals of Statistics*.

Recap: Post-selection Inference

For each $\mathrm{M} \subseteq \{1, 2, \ldots, p\}$,

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

Want a valid CI for $\beta_{\widehat{j}\cdot\widehat{\mathrm{M}}}$, a coordinate of $\beta_{\widehat{\mathrm{M}}}$:

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

irrespective of how $\widehat{\mathrm{M}}$ with size $\leq k$ and $\widehat{j} \in \widehat{\mathrm{M}}$ are chosen based on data.

Literature Review

Buehler and Feddersen (1963); Olshen (1973); Sen (1979); Rencher and Pun (1980); Freedman (1983); Sen and Saleh (1987); Dijkstra and Veldkamp (1988); Hurvich and Tsai (1990); Potscher (1991); Pfeiffer, Redd and Carroll (2017).

Cox (1965); Kabaila (1998); Hjort and Claeskens (2003); Claeskens and Carroll (2007); Berk et al. (2013); Lee, Sun, Sun and Taylor (2016); Tibshirani, Taylor, Lockhart and Tibshirani (2016); Bachoc, Preinerstorfer and Steinberger (2019); Rinaldo, Wasserman, G'Sell and Lei (2019).

Roadmap

Data Snooping: Effects and Examples

- Effects illustrated with stepwise selection.
- Data snooping in textbooks and practice.

Formulation of the Problem

- The Problem & literature review for covariate selection.

Solution for Covariate Selection

- Key contributions
- Simulations & main components of the theory.

Example and Conclusions

A Guiding Principle

Simultaneous Inference FWER Control $\Rightarrow$ Post-selection Inference

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

- Simultaneity implies valid CIs for arbitrary selection $\widehat{\mathrm{M}}$
- Simultaneity implies *infinite* revisions of a selection.
- Simultaneity also guarantees validity if multiple models are reported.

A Key Result

Simultaneous Inference FWER Control $\Longleftrightarrow$ Post-selection Inference

$$\mathbb{P}\left(\bigcap_{|\mathrm{M}|\leq k,\; j\in\mathrm{M}}\left\{\beta_{j\cdot\mathrm{M}} \in \widehat{\mathbf{CI}}_{j\cdot\mathrm{M}}\right\}\right) = \inf_{\substack{|\widehat{\mathrm{M}}|\leq k,\\ \hat{j}\in\widehat{\mathrm{M}}}} \mathbb{P}\left(\beta_{\hat{j}\cdot\widehat{\mathrm{M}}} \in \widehat{\mathbf{CI}}_{\hat{j}\cdot\widehat{\mathrm{M}}}\right)$$

Theorem: Simultaneous inference is *necessary* for valid Post-selection inference.

K et al. (2019) Valid Post-selection Inference in Model-free Linear Regression. *Annals of Statistics (Forthcoming)*.

Solution 1: Uniform Adjustment

The classical interval for $\beta_{j\cdot\mathrm{M}}$ is

t-score

$$\left\{\theta : \left|\boxed{\frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}}-\theta)}{\widehat{\sigma}_{j\cdot\mathrm{M}}}}\right| \le z_{\alpha/2}\right\}.$$

For *simultaneity*, inflate the confidence regions:

$$\widehat{\mathrm{CI}}^{\mathrm{PoSI}}_{j\cdot\mathrm{M}} := \left\{\theta : \left|\frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}}-\theta)}{\widehat{\sigma}_{j\cdot\mathrm{M}}}\right| \le K_\alpha\right\},$$

$K_\alpha = (1-\alpha)$ quantile of $\displaystyle\max_{|\mathrm{M}|\le k,\, j\in\mathrm{M}} \left|\frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}}\right|$

Our Contributions

We show

- in an assumption-lean setting,
- for independent and weakly dependent obs.,
- for $p \gg n$ and maximal model size $k = k_n$,
- for fixed or random covariates,

K_α **can be estimated using bootstrap.**

In the worst case,

$$K_\alpha \asymp \sqrt{k\,\mathrm{Log}(p/k)}. \qquad (\mathrm{Log}(x) = 1 + \log x)$$

Berk et al. (2013): **Homoscedastic Gaussian** response, **fixed X**.
Bachoc et al (2019): assumption-lean but **fixed X** and **fixed *p***.

Simulation Examples:

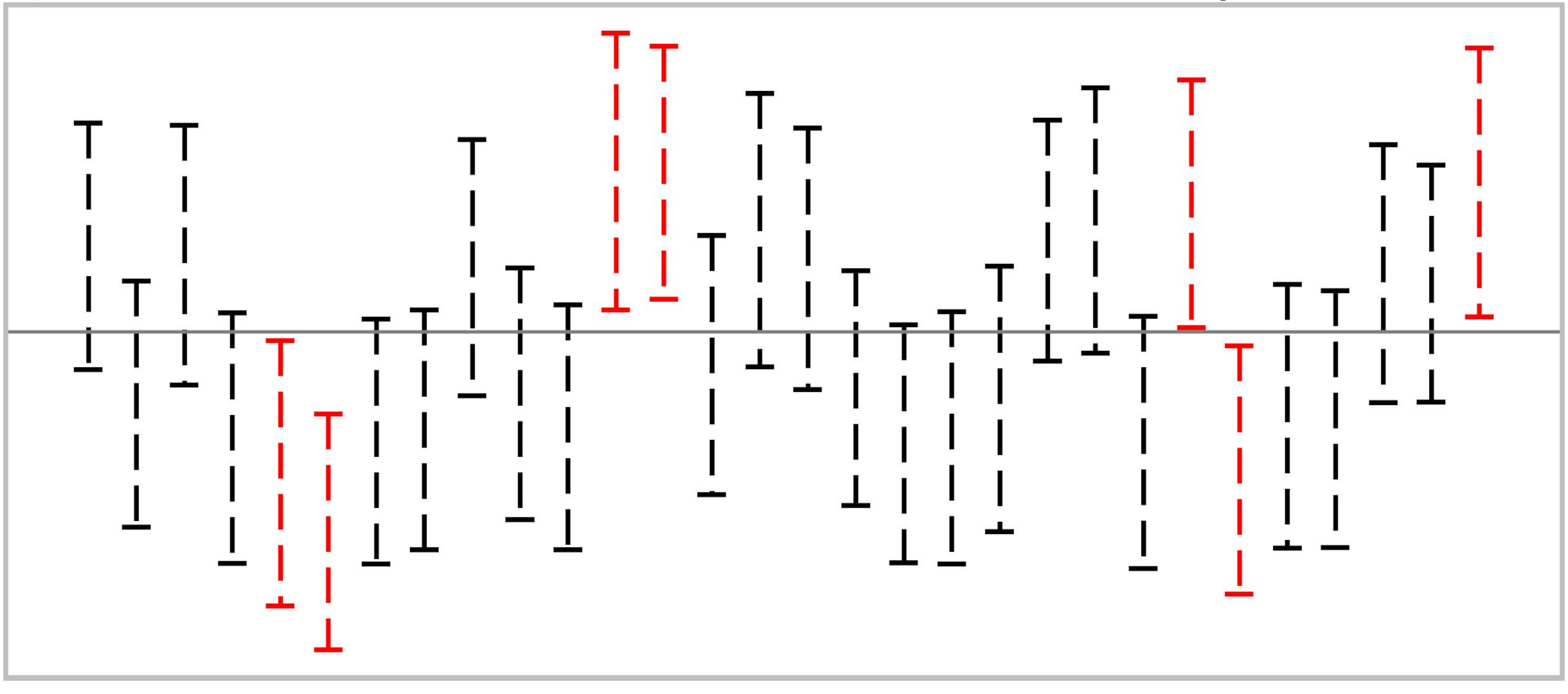
Revisiting Stepwise Selection

Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1})$ $\Rightarrow$ 500 observations

$X_{\hat{j}}$ most correlated with Y $\Rightarrow$ 95% CI for slope

p = **5** Unadjusted: **76.9%**

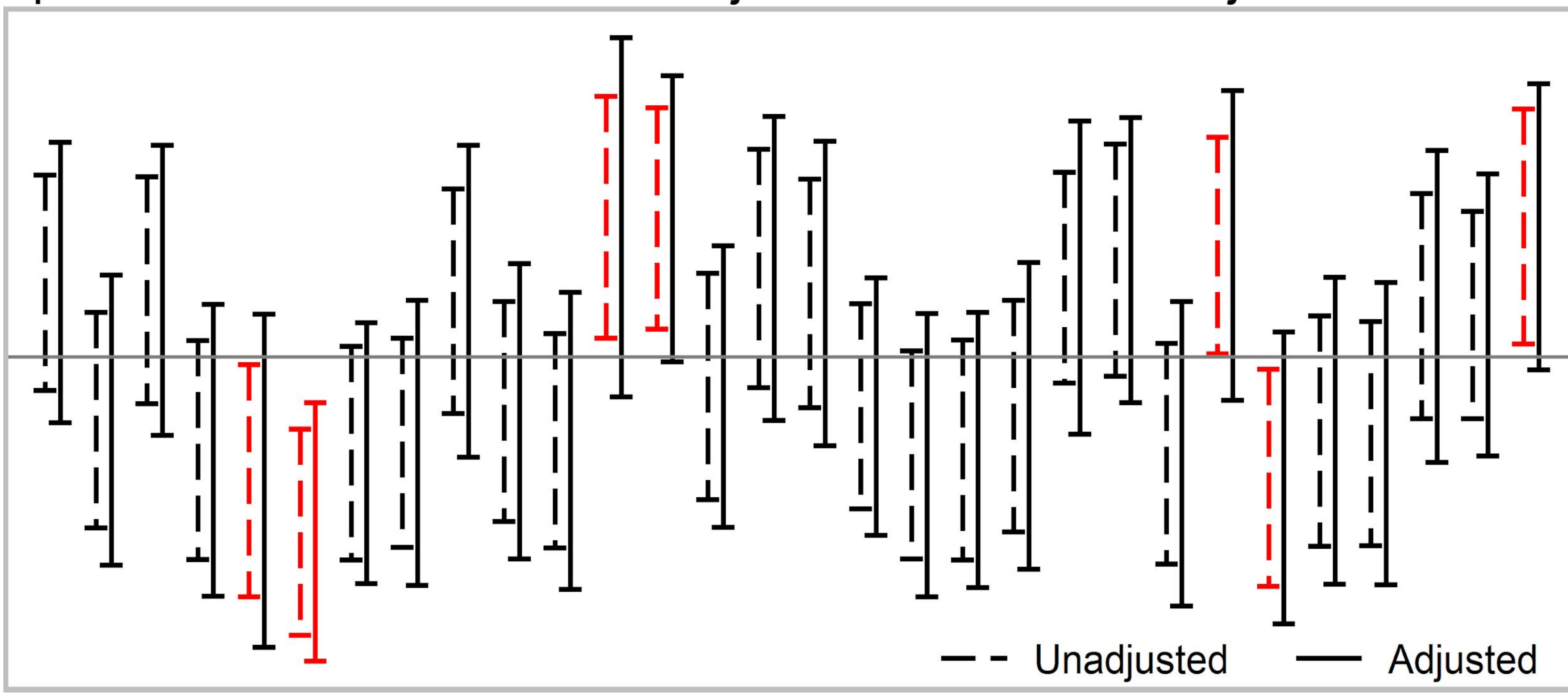


Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1})$ $\Rightarrow$ 500 observations

$X_{\hat{j}}$ most correlated with Y $\Rightarrow$ 95% CI for slope

p = **5** Unadjusted: **76.9%** Adjusted: **95.2%**

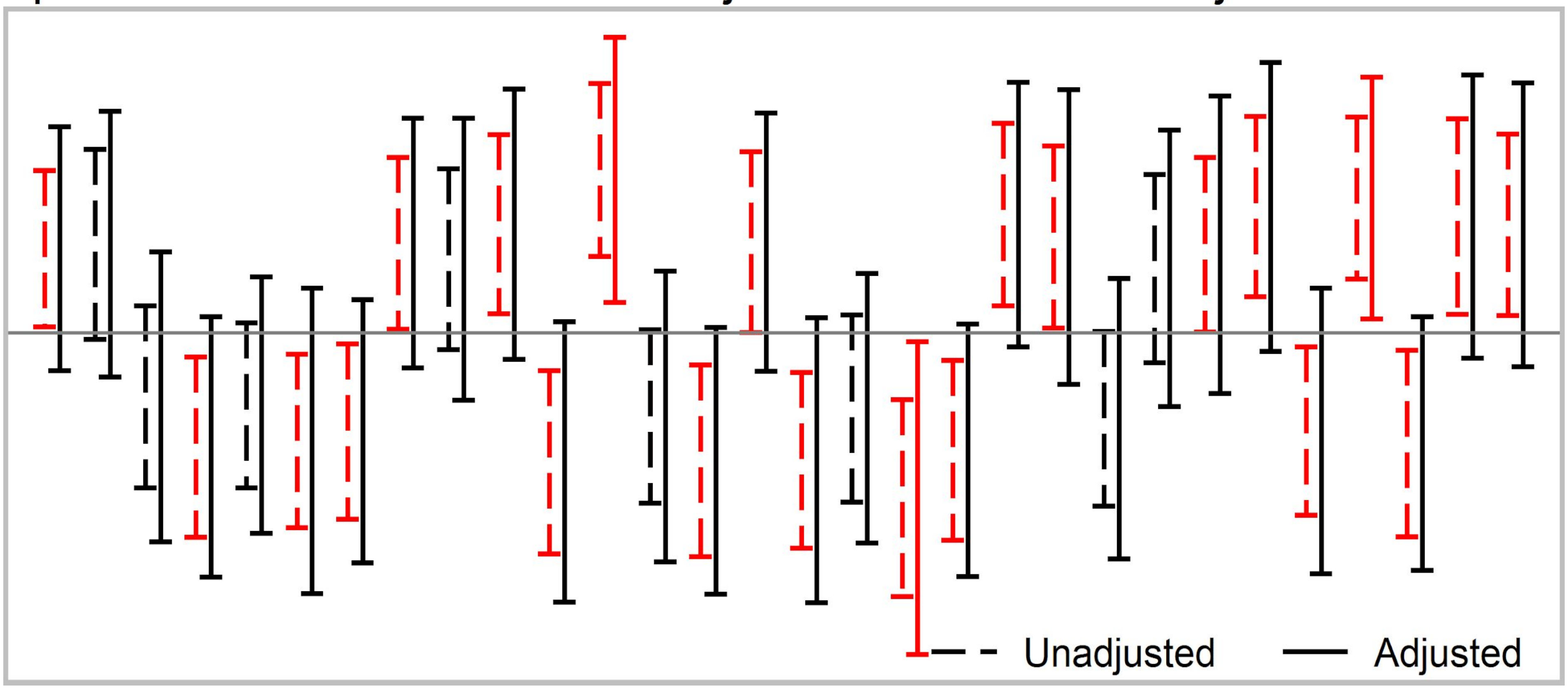


Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1})$ $\Rightarrow$ 500 observations

$X_{\hat{j}}$ most correlated with Y $\Rightarrow$ 95% CI for slope

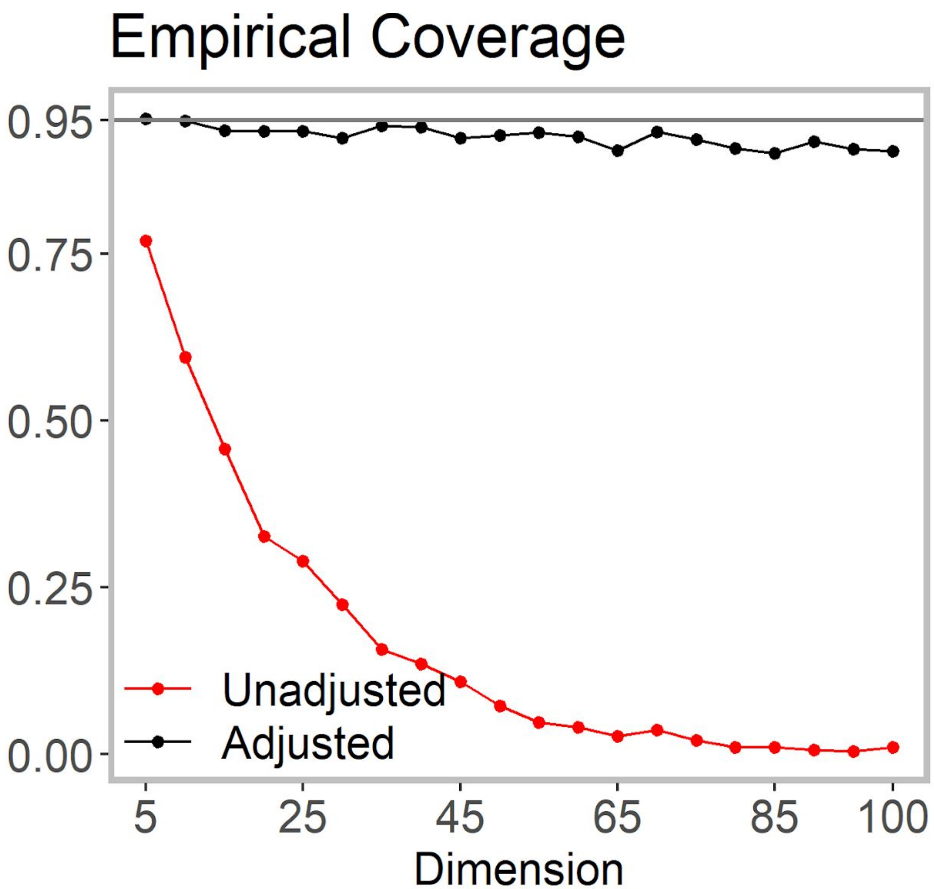
p = **20** Unadjusted: **32.6%** Adjusted: **93.4%**

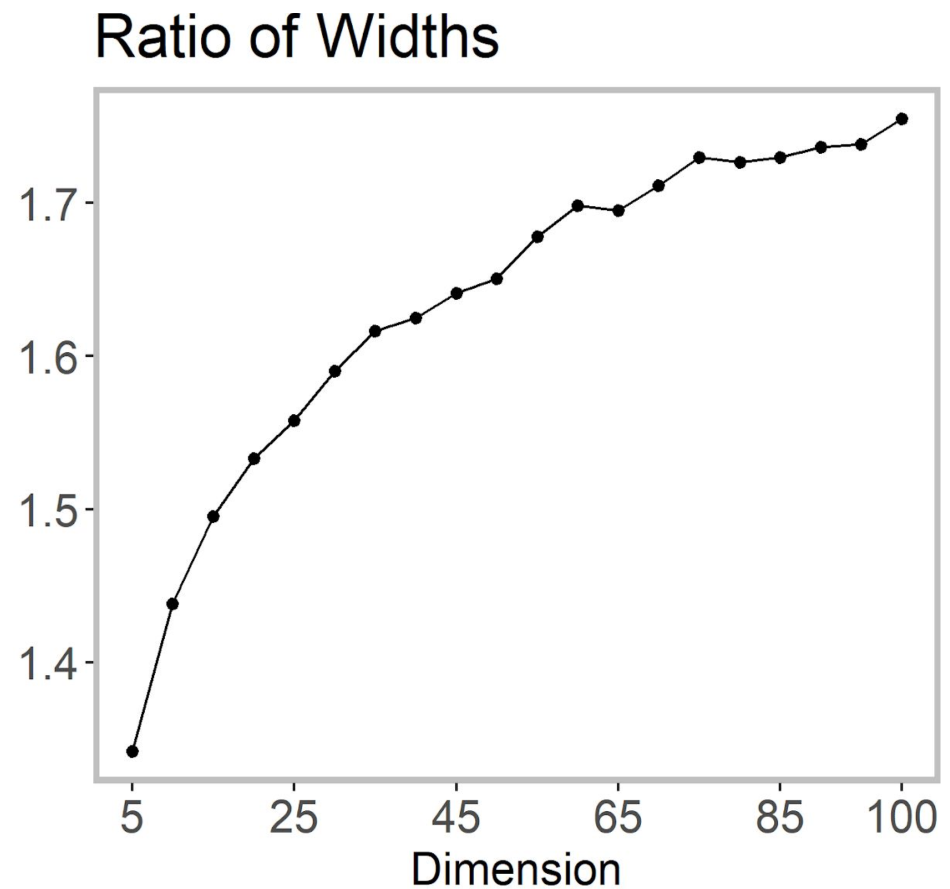


Effects of Data Snooping: Stepwise Selection

$(X, Y) \sim N(0, I_{p+1}) \quad \Rightarrow \quad$ 500 observations

$X_{\hat{j}}$ most correlated with $Y \quad \Rightarrow \quad$ 95% CI for slope





Key Steps in the Proof

Uniform Linear Representation

For independent, sub-Gaussian data $(X_i, Y_i), 1 \leq i \leq n$

$$\max_{\substack{|\mathrm{M}|\leq k,\\ j\in\mathrm{M}}} \left| \frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{j\cdot\mathrm{M}}(X_i, Y_i) \right| = O_p\left(\frac{k\,\mathrm{Log}(p/k)}{\sqrt{n}} \right).$$

- ➢ Doesn't require any parametric model assumptions.
- ➢ A finite sample result. Allows for diverging p, k.
- ➢ Extends beyond independent & sub-Gaussian data.

K et al. (2018) A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference. arXiv:1802.05801

Three Line Proof

$$\max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}} \left| \frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{j\cdot\mathrm{M}}(X_i, Y_i) \right| = O_p\left(\frac{k\,\mathrm{Log}(p/k)}{\sqrt{n}} \right).$$

implies

$$\max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}} \left| \frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}} \right| \xrightarrow[\text{by triangle ineq.}]{\text{Close in Probability}} \max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{j\cdot\mathrm{M}}(X_i, Y_i) \right|$$

Three Line Proof

$$\max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}}\left|\frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}}-\beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{j\cdot\mathrm{M}}(X_i,Y_i)\right| = O_p\left(\frac{k\,\mathrm{Log}(p/k)}{\sqrt{n}}\right).$$

implies

$$\max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}}\left|\frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}}-\beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}}\right|$$

Close in Probability
by triangle ineq.

$$\max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{j\cdot\mathrm{M}}(X_i,Y_i)\right|$$

Close in
Distribution
by HDCLT

$$\max_{|\mathrm{M}|\le k,\ j\in\mathrm{M}}|G_{j\cdot\mathrm{M}}|$$

$$\boxed{(k\,\mathrm{Log}(p/k))^5 = o(n)}$$

K et al. (2018) High-dimensional CLT: Improvements, Non-uniform Extensions and Large Deviations. arXiv:1806.06153

Three Line Proof

$$\max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}} \left| \frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{j\cdot\mathrm{M}}(X_i, Y_i) \right| = O_p\left(\frac{k\,\mathrm{Log}(p/k)}{\sqrt{n}}\right).$$

implies

$$\max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}} \left| \frac{\sqrt{n}(\widehat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}})}{\widehat{\sigma}_{j\cdot\mathrm{M}}} \right| \xrightarrow[\text{by triangle ineq.}]{\text{Close in Probability}} \max_{\substack{|\mathrm{M}|\le k,\\ j\in\mathrm{M}}} \left| \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{j\cdot\mathrm{M}}(X_i, Y_i) \right|$$

Close in
Distribution
by triangle ineq.

Close in
Distribution
by HDCLT

$$\max_{|\mathrm{M}|\le k,\; j\in\mathrm{M}} \left|G_{j\cdot\mathrm{M}}\right|$$

Justifies bootstrap!

Roadmap

Data Snooping: Effects and Examples

- Effects illustrated with stepwise selection.
- Data snooping in textbooks and practice.

Formulation of the Problem

- The Problem & literature review for covariate selection.

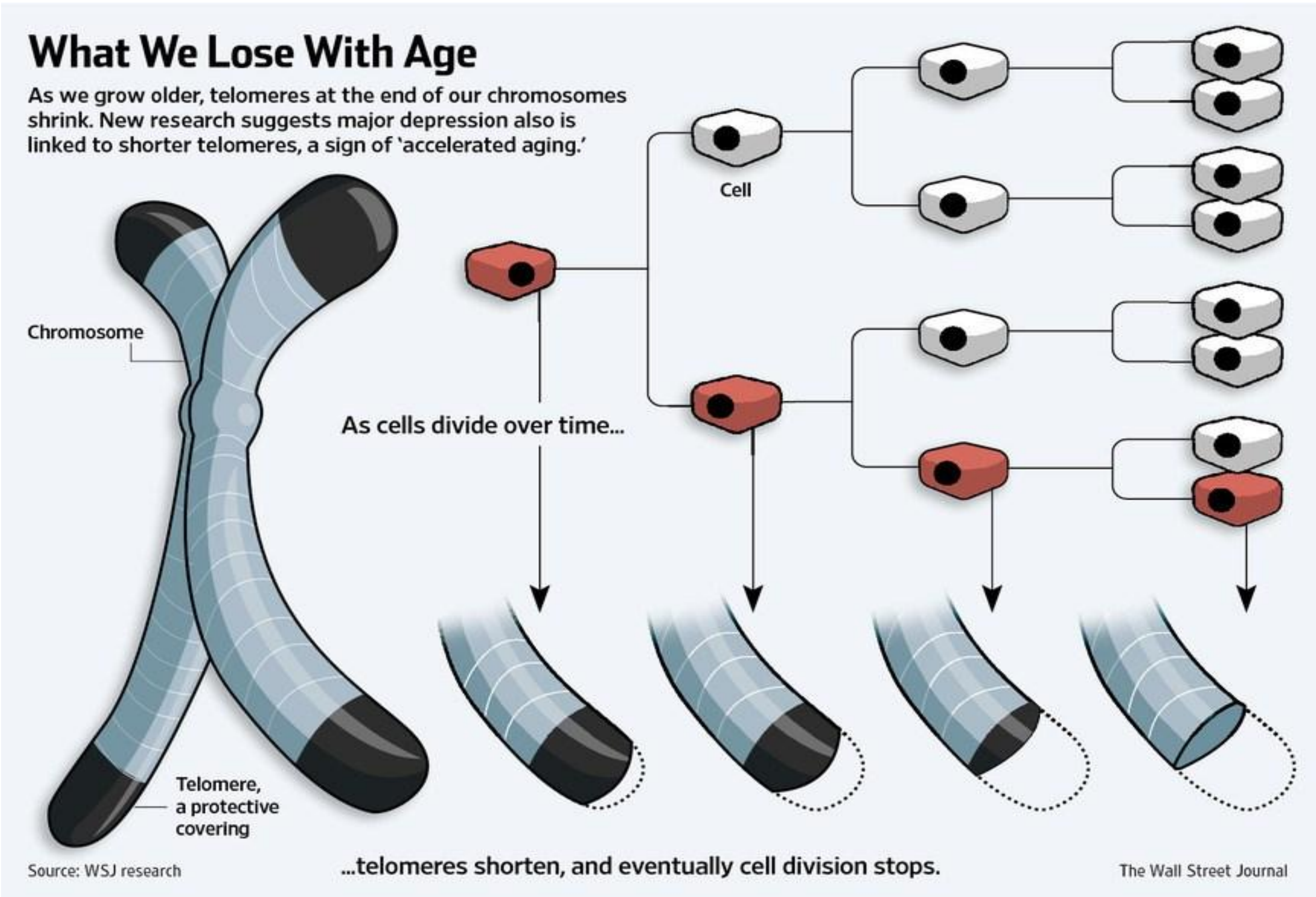
Solution for Covariate Selection

- Key contributions
- Simulations & main components of the theory.

Example and Conclusions

- Real Data Example, Extensions & Summary.

Telomere Length Analysis



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."

Telomere Length Analysis

TL inheritance patterns based on 246 families.

- *Dependent Variable*: **MTL** (Mean telomere length)
- *Child Variables*:
  - **Sex**
  - **Age**
- *Parental Variables*:
  - **mMTL** (mother MTL)
  - **fMTL** (father MTL)
  - **MAC** (mother's age at conception)
  - **PAC** (father's age at conception)

(Additionally, 15 interaction variables were considered.)

Adjusted Inference: Telomere Length Analysis

Covariate	Unadjusted	Adjusted
AGE	✓	✗
mMTL	✓	✓
fMTL	✓	✓
MAC	✓	✗
PAC	✗	✗

✓ : Significant at 5% level

✗ : Insignificant at 5% level

Summary and Conclusions

- ❖ Data snooping contributes to replicability crisis.
- ❖ Inference is possible after data snooping.

Classical Framework	New Framework
Fix the test & model	Fix a universe
Collect the data	Collect the data

- ❖ The framework allows for
  - ➢ Misspecified models; Random covariates;
  - ➢ Dependent data; High-dimensional features;
  - ➢ Variable Transformations;
  - ➢ M-estimators: logistic/Poisson/Quantile/Cox.

References

Kuchibhotla A., Brown L., Buja A., Cai J., George E., Zhao L. (2019)
Valid Post-selection Inference in Model-free Linear Regression.
Annals of Statistics (Forthcoming).

Kuchibhotla A., Brown L., Buja A., George E., Zhao L. (2018)
A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference. *arXiv:1802.05801*

Kuchibhotla A. (2018)
Deterministic Inequalities for Smooth M-estimators. *arXiv:1809.05172*

Kuchibhotla A., Mukherjee S., and Banerjee D. (2018)
High-dimensional CLT: Improvements, Non-uniform Extensions and Large Deviations. *arXiv:1806.06153*

Kuchibhotla A., Brown L., Buja A., Cai J. (2019)
All of Linear Regression. *arXiv:1910.06386*

Thank you for your attention

Post-selection for Transformations

For each $\mathrm{g} \in \mathcal{G} \subseteq L_2(Y)$,

$$\beta_{\mathrm{g}} := \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \mathbb{E}[(\mathrm{g}(Y) - X^\top \theta)^2].$$

Want a valid CI for $\beta_{1 \cdot \widehat{\mathrm{g}}}$, a coordinate of $\beta_{\widehat{\mathrm{g}}}$

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

irrespective of how $\widehat{\mathrm{g}} \in \mathcal{G}$ is chosen based on data.

Solution for Transformations

Inflate the classical intervals,

$$\left\{\theta \in \mathbb{R} : \left|\frac{\sqrt{n}(\widehat{\beta}_{1\cdot \mathrm{g}} - \theta)}{\widehat{\sigma}_{1\cdot \mathrm{g}}}\right| \le K_\alpha\right\},$$

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

$$\max_{\mathrm{g}\in\mathcal{G}} \left|\frac{\sqrt{n}(\widehat{\beta}_{1\cdot \mathrm{g}} - \beta_{1\cdot \mathrm{g}})}{\widehat{\sigma}_{1\cdot \mathrm{g}}}\right|,$$

Bootstrap applies with validity for “nice” function classes $\mathcal{G}$, for example, Box-Cox family.

Transformations: Boston Housing Data

This dataset has 506 census tracts with 13 features.

- *Dependent Variable*: **MV**, Median value of house.
- *Covariate of Interest*: **NOX**, Nitrogen Oxide Conc.
- *Structural Variables*: **RM** (No. of rms), **AGE** (% of homes bfr 1940).
- *Neighborhood Variables*: **CRIM** (Crime rate), **ZN** (% of res. land zoned for lots > 25K ft^2), **INDUS** (% non-retail business acres per twn), **RIVER** (Charles river dummy), **TAX** (Property tax rate), **PTRATIO** (Pupil-teacher ratio), **B** (Racial diversity), **LSTAT** (% of lwr socio-econ. status of population).
- *Accessibility Variables*: **DIS** (Distance to Employment Ctr.), **RAD** (Distance to Radial Highway).

Transformations: Boston Housing Data

Harrison and Rubinfeld (1978)★ write

to determine the best fitting functional form. Comparing models with either median value of owner-occupied homes (MV) or Log(MV) as the dependent variable, we found that the semilog version provided a slightly better fit. Using Log(MV) as the dependent variable, we concentrated on estimating a nonlinear term in NOX; i.e., we included NOX^p in the equation, where p is an unknown

The statistical fit in the equation was best when p was set equal to 2.0,

★H & R (1978) Hedonic housing prices and the demand for clean air.

Adjusted Inference: Boston Housing

Covariate	Unadjusted	Adjusted	Covariate	Unadjusted	Adjusted
NOX^2	✓	✓	**TAX**	✓	✓
RM	✓	✓	**PTRATIO**	✓	✓
AGE	✗	✗	**B**	✓	✗
CRIM	✓	✓	**LSTAT**	✓	✓
ZN	✗	✗	**DIS**	✓	✓
INDUS	✗	✗	**RAD**	✓	✓
RIVER	✓	✗			

✓ : Significant at 5% level

✗ : Insignificant at 5% level

Implications of PoSI for Applications

Similar to Boston housing data and TL data,

How do conclusions in applied data analysis change when exploration is accounted for?

Joint work with Junhui Cai and Linda Zhao.

High-dimensional CLT

Anderson, Hall and Titterington (1998, JSPI):

Let $\mathscr{R}$ be the class of all rectangles in $\mathbb{R}^p$.

A proof will be given in Section 3.2. It follows from this result and the Theorem that if (2.1), (2.6) and (2.7) hold,

$$\sup_{B \in \mathscr{R}} \left| P(S \in B) - \int_B \phi(x)\,\mathrm{d}x \right| = \mathrm{O}\{n^{-1/2}(\log\, p)^{3/2}\}.$$

Maximal Inequalities

- High-dimensional rates are sensitive to tails.

- What is the order of

$$\mathbb{E}\left[\max_{1\le j\le d}\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right|\right], \quad \text{subject to}$$

$$\max_{1\le j\le d} \mathrm{Var}(X_{1,j}) \le A^2 \quad \text{and} \quad \mathbb{E}\left[\|X_1\|_\infty^q\right] \le B^q?$$

Joint work with Somabha Mukherjee and Sagnik Nandy.

Multiple Testing under Dependence

- Multiple testing often requires independence.
- FWER is possible under arbitrary dependence.
- What about FDR control?
- How does BH procedure behave?

Arun Kumar Kuchibhotla

University of Pennsylvania

Webpage: https://arun-kuchibhotla.github.io/

Email: arunku@upenn.edu

Case Study 3: Model Building

Modeling Home Prices Using Realtor Data

Iain Pardoe
Lundquist College of Business, University of Oregon

Journal of Statistics Education
2008

- 76 Oregon homes and 12 features.
- Try a linear model with 12 predictors "as is".
- Residuals imply non-linearity. $\text{Age} \rightarrow \text{Age}^2$.
- Bath and Bed also have high p-values, so add an interaction $\text{Bath} \times \text{Bed}$ to the model.
- Price is skewed suggesting a log-transformation.

Case Study 3: Transformations

In the context of curve fitting to bivariate data, Stine and Foster (2014)★ on page 515, write

> "Picking a transformation requires practice, and you may need to try several to find one that is interpretable and captures the pattern in the data."

★Stine and Foster, Statistics for Business: Decision Making and Analysis.

Extensions

- The solution applies to most other estimation problems.

- Examples include logistic/Poisson regression, quantile regression and Cox regression.

- Solution also applies to transformation of variables.

K (2018) Deterministic Inequalities for Smooth M-estimators, arXiv:1809.05172