Valid Post-selection Inference in Model-free Linear Regression[a]

Arun Kumar Kuchibhotla

14 Decemeber, 2019

The Wharton School,
University of Pennsylvania.

[a] Joint work with Larry Brown, Andreas Buja, Junhui Cai, Ed George and Linda Zhao

Table of contents

Invalidity of Classical Inference

Data snooping is an integral part of data analysis.

For regression analysis, variable selection is a result of such snooping.

Classical inference after such variable selection can be misleading.

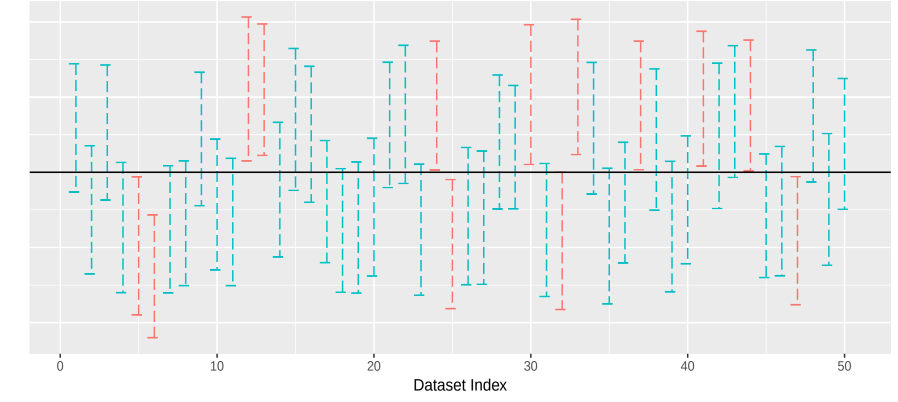
Example: Invalidity of classical inference under selection

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

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

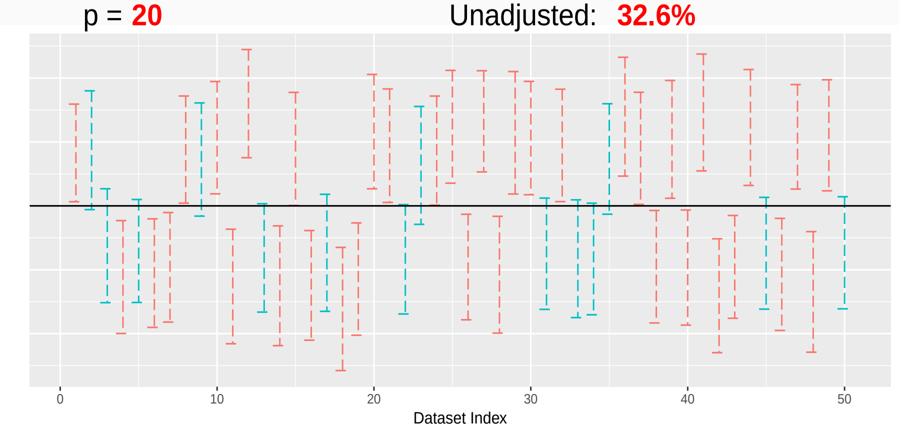


Example: Invalidity of classical inference under selection

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

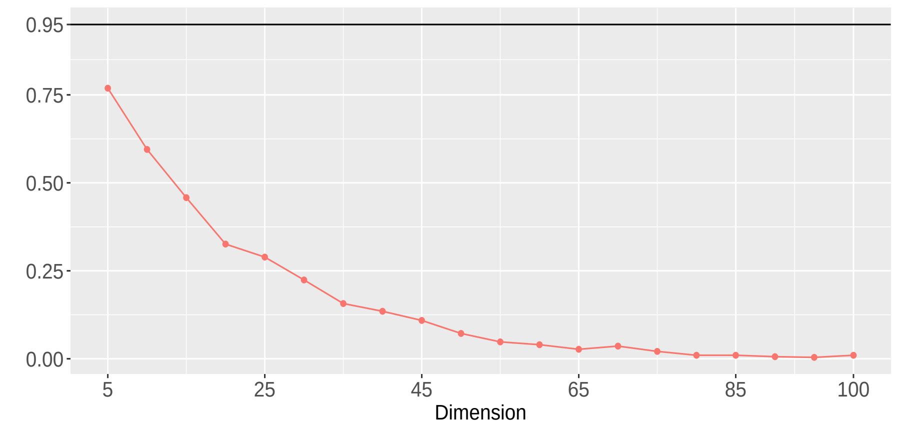


Example: Invalidity of classical inference under selection

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.

Forward selection is a conventional variable selection strategy and is very commonly taught in basic courses.

Formulation of the Problem

There are p hypotheses to start with

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

Equivalently,

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

where

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

There are p hypotheses to start with

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

Equivalently,

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

where

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

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

There are p hypotheses to start with

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

Equivalently,

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

where

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

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

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

Classical (invalid) test:

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

The General PoSI Problem

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

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

Fix k: $1 \leq k \leq p$. Construct confidence regions $\widehat{\mathrm{CI}}_{\widehat{j}\cdot\widehat{\mathrm{M}}}$ such that

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

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

Simultaneous Inference $\Rightarrow$ **Post-selection Inference**
(FWER Control)

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

Theorem: **FWER control is *necessary* for valid PoSI.**

Three Solutions

A (Very) Simple Solution

Apply Bonferroni procedure.

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

How many elements in the sum?

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

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

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

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

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

Accounts for dependence.

Disadvantage of these Solutions

Bonferroni Solution:

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

PoSI Solution:

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

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

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

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

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

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

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

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

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

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

Idea of the Proof

For any model M, the OLS estimator is

$$\widehat{\beta}_{\mathrm{M}} := \operatorname*{argmin}_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_{i,\mathrm{M}}^\top \theta)^2 \quad \equiv \quad \frac{1}{n}\sum_{i=1}^{n} X_{i,\mathrm{M}}\left(Y_i - X_{i,\mathrm{M}}^\top \widehat{\beta}_{\mathrm{M}}\right) = 0.$$

Adding and subtracting β_{M} leads to

$$\begin{aligned}\widehat{\Sigma}_{\mathrm{M}}(\widehat{\beta}_{\mathrm{M}} - \beta_{\mathrm{M}}) &= \frac{1}{n}\sum_{i=1}^{n} X_{i,\mathrm{M}}(Y_i - X_{i,\mathrm{M}}^\top \beta_{\mathrm{M}}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{X_{i,\mathrm{M}}(Y_i - X_{i,\mathrm{M}}^\top \beta_{\mathrm{M}}) - \mathbb{E}[X_{i,\mathrm{M}}(Y_i - X_{i,\mathrm{M}}^\top \beta_{\mathrm{M}})]\right\},\end{aligned}$$

Thus,

$$\|\widehat{\Sigma}_{\mathrm{M}}(\widehat{\beta}_{\mathrm{M}} - \beta_{\mathrm{M}})\|_\infty \quad \le \quad \left\|\frac{1}{n}\sum_{i=1}^{n}(X_i Y_i - \mathbb{E}[X_i Y_i])\right\|_\infty + \|\!|\widehat{\Sigma} - \Sigma|\!\|_\infty \|\beta_{\mathrm{M}}\|_1,$$

for all models $\mathrm{M} \subseteq \{1, 2, \ldots, p\}$.

Further if the observations are *independent* and $\|X_i\|_\infty$ has finite second moment, then for any random model $\widehat{\mathrm{M}}$ with $|\widehat{\mathrm{M}}| = o_p(\sqrt{n/\log p})$,

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

where for any model M,

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

- These confidence regions are not rectangles but are parallelepipeds.
- It is fairly trivial to project these regions to get confidence intervals for $\beta_{j\cdot\widehat{\mathrm{M}}}$.
- Calculating $\widehat{\mathrm{CI}}_{\widehat{\mathrm{M}}}$ only requires computing $\widehat{\beta}_{\widehat{\mathrm{M}}}$, $C_{xx}(\alpha)$ and $C_{xy}(\alpha)$.
- Computational cost: $O(p^2)$ times the number of bootstrap samples.

Theoretical and Numerical Comparison

Comparison of Volumes

Reference	**Leb**$(\widehat{\mathrm{CI}}_{\widehat{\mathrm{M}}})$	Design
Kuchibhotla et al. (2019)	$(\log p/n)^{\lvert\widehat{\mathrm{M}}\rvert/2}$	fixed
	$(\lvert\widehat{\mathrm{M}}\rvert \log p/n)^{\lvert\widehat{\mathrm{M}}\rvert/2}$	random
Berk et al. (2013) Bachoc et al. (2019) Kuchibhotla et al. (2018)	$(k \log(ep/k)/n)^{\lvert\widehat{\mathrm{M}}\rvert/2}$	fixed/random
Taylor and Co. (2016+)	Infinite	fixed/random

Table 1: Volumes of Different PoSI Regions.

Simulations

Setting:

$$\mathbf{Y} \;=\; \mathbf{X}\beta_0 + \xi, \quad \text{where} \quad \beta_0 = \mathbf{0}_p, \quad \xi \sim N(0, I_n).$$

We consider fixed covariates with following designs:

- **Orthogonal design:**

$$\frac{\mathbf{X}^\top \mathbf{X}}{n} = \widehat{\Sigma} = I_p.$$

- **Equicorrelation design:**

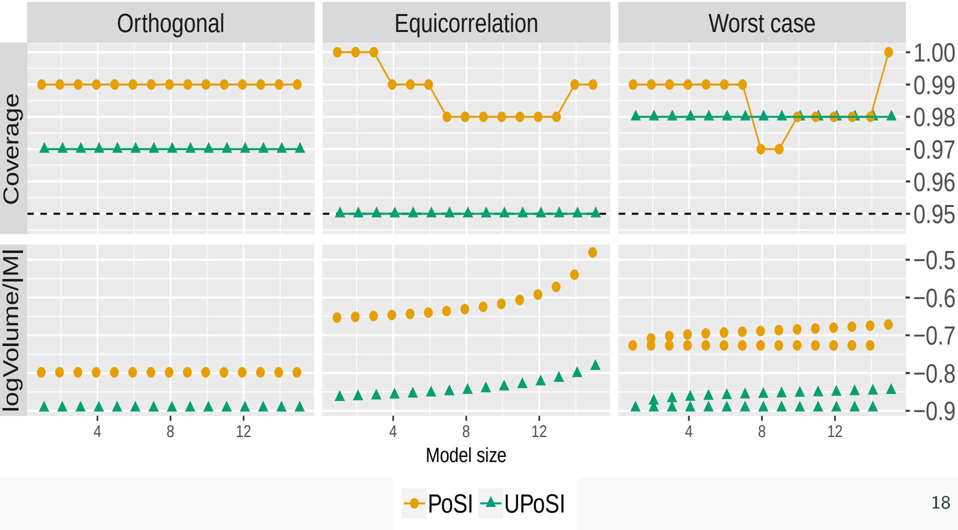
$$\widehat{\Sigma} = I_p + \alpha \mathbf{1}_p \mathbf{1}_p^\top \quad \text{with} \quad \alpha = -\frac{1}{(p+2)}.$$

- **Wors-case design:**

$$\widehat{\Sigma} \;=\; \begin{bmatrix} I_{p-1} & c\mathbf{1}_{p-1} \\ \mathbf{0}_{p-1}^\top & \sqrt{1-(p-1)c^2} \end{bmatrix}, \quad \text{with} \quad c^2 = \frac{1}{2(p-1)}.$$

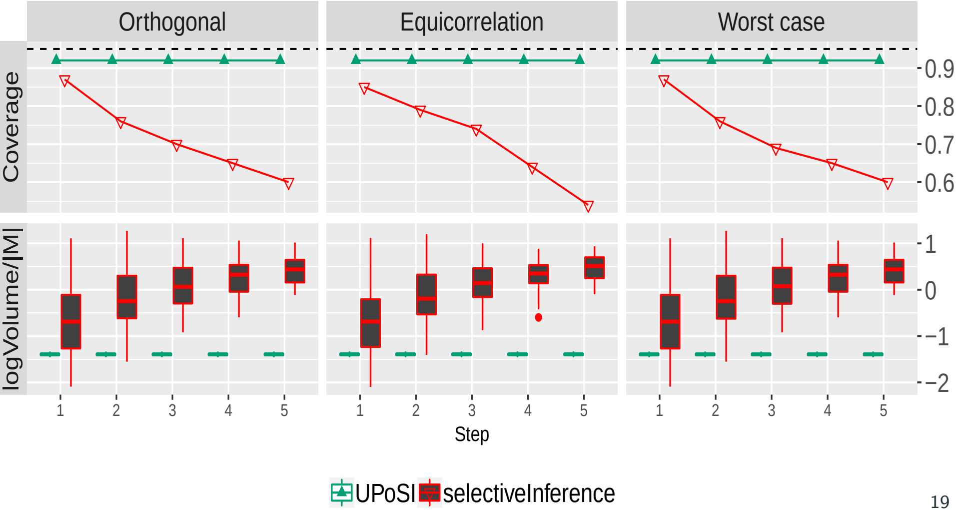
Comparison with PoSI: max-t statistic

200 observations and 15 covariates.



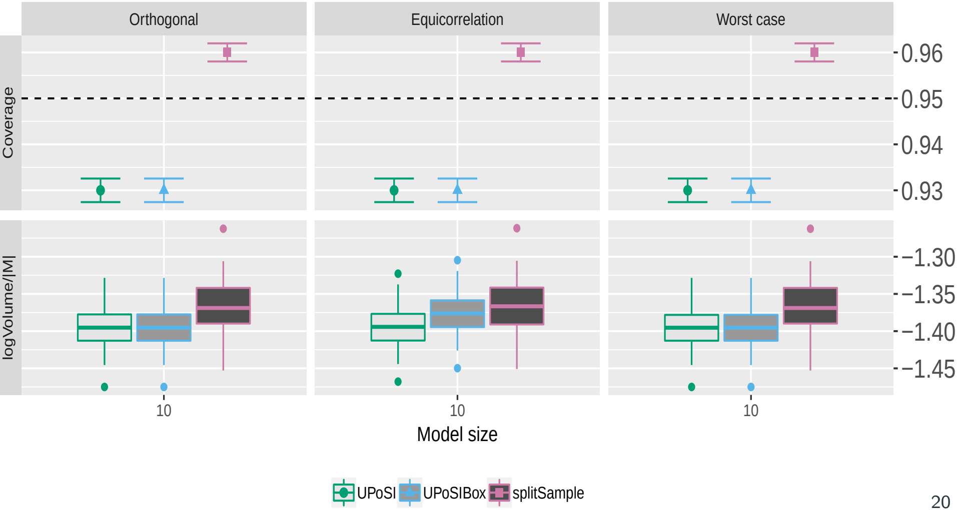
Comparison with Selective Inference: Forward Stepwise

1000 observations and 500 covariates.



Comparison with Sample Splitting: Forward Stepwise

1000 observations and 500 covariates.



Conclusions

Provided a computationally efficient post-selection inference for linear regression.

The proposed regions have better volume properties than existing alternatives.

Does not require any of the classical linear modeling assumptions.

Works for dependent observations as well.

Crucial ingredient: Bootstrap for estimating quantiles.

Reference: Kuchibhotla et al. (2019) Valid Post-selection Inference in Model-free Linear Regression, Annals of Statistics. Forthcoming.