Post-Selection Inference and Misspecification[1]

Andreas Buja & Arun K. Kuchibhotla

Department of Statistics
University of Pennsylvania

June 4, 2018

[1]Joint work with "Larry's Group" at Wharton, including Larry Brown, Edward George and Linda Zhao.

Outline

In Memory of

LARRY BROWN †

February 21, 2018

Outline

Machine Learning vs Statistics

- **Some classical machine learning themes:**
  - Prediction: click-through, consumer choices, investment returns, ...
  - Classification: images, speech, text, ...
  - Online decision making

$\Longrightarrow$ **Construction of data-driven black boxes, automation for Technology**

- **A classical statistics theme:**
  - SEs, tests, p-values, CIs (2 meanings), posteriors, ... for

STATISTICAL INFERENCE

$\Longrightarrow$ **Knowledge acquisition by humans for Science**

A Crisis in the Sciences: Irreproducibility

- Indicators of a crisis:
  - 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"
  - Simmons, Nelson, Simonsohn (2011, Psychol.Sci):
    "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant"

$\Longrightarrow$ **Irreproducibility of Empirical Findings**

- Many potential causes – two major ones:
  - Institutional: Publication bias, "file drawer problem"
  - Methodological: Statistical biases, **"researcher degrees of freedom"**

Irreproducibility: Methodol. Factor 1 – Selection

- A statistical bias is due to lack of accounting for
  selection of variables, transforms, scales, subsets, weights, ...
- **Regressor/model selection** (our focus) is on several levels:
  - formal selection: all subset (C_p, AIC, BIC,...), stepwise (F), lasso,...
  - informal selection: diagnostics for GoF, influence, collinearity,...
  - post hoc selection: “Effect size is too small, the variable too costly.”
- Suspicions and Criticisms:
  - All three modes of selection are (should be) used.
  - More thorough data analysis $\Longrightarrow$ More spurious results
  - Not a solution: Post-selection inference for “adaptive Lasso”, say.
    Empirical researchers do not write contracts with themselves to commit **a priori** to one formal selection method and nothing else.

The "PoSI" Solution to Selection: FWER Control

- PoSI Procedure — general version:
  - Define a **universe** $\mathcal{M}$ of models M you might ever consider/select: outcomes (Y), regressors (X), their transforms ($f(X), g(Y)$), ...
  - Define the universe of all tests you might ever perform in these models, typically for regression coeffs $\beta_{j,M}$ (j'th coeff in model M).
  - Consider the **minimum of the p-values** for all these tests: Obtain its 0.05 quantile $\alpha_{0.05}$ for FWER adjustment.
  - Now freely examine your data and select models $\hat{M} \in \mathcal{M}$, reconsider, re-select, re-reconsider, ... but compare all p-values against $\alpha_{0.05}$, not 0.05, for **0.05 $\mathcal{M}$-FWER control**.
- Cost-Benefit Analysis:
  - Cost: Huge computation upfront — adjustment for millions of tests
  - Benefits: **Solution to the circularity problem** — select model $\hat{M}$, don't like it, select $\hat{M}'$, don't like it, ... PoSI inference remains valid.

Irreproducibility: Methodol. Factor 2 – Misspecification

- Models are approximations, not generative truths.
  $\Longrightarrow$ Consequences!
- What is the target $\beta_{j,M}$ of $\hat{\beta}_{j,M}$? Stay tuned.
- Model bias interacts with regressor distributions to cause model-trusting SEs to be off, sometimes too small by a factor of 2.

$$V[\hat{\beta}] \;=\; E[V[\hat{\beta}|X]] + V[E[\hat{\beta}|X]]$$

- Do not condition on the regressors; do not treat them as fixed!
- Use model-robust standard errors, for example, from the *x*-*y* pairs or multiplier bootstraps, **not** the residual bootstrap!

Wanted: PoSI Protection under Misspecification!

Up next: Asymptotic theory for Regressor Selection

Outline

Linear Regression: A Simple Question

Suppose $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ are n independent random vectors and the least squares linear regression estimator $\hat{\beta}_n$ is computed, that is,

$$\hat{\beta}_n := \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^\top\theta)^2,$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

is computed assuming the matrix above is invertible.

Problem

What is $\hat{\beta}_n$ estimating? Can there be a justification for this without the usual Gauss-Markov assumptions? Is independence necessary?

Note that random vectors can be non-identically distributed.

Linear Regression

- From definition, $\hat{\beta}_n$ is a smooth (non-linear) function, $G(\cdot,\cdot)$, of two averages:

$$\hat{\beta}_n = G\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top,\ \frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right).$$

- If the random vectors (X_i, Y_i) are such that these averages converge to their expectations, then by Slutsky's theorem

$$\hat{\beta}_n - \beta_n = o_p(1),$$

  where

$$\beta_n := G\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^\top\right],\ \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i Y_i\right]\right).$$

- Hence there exists a target of estimation under "minimal" assumptions that require neither linearity nor homoscedastic Gaussian errors.

Inference under Misspecification

- Recall $\hat{\beta}_n$ estimates the best linear projection in the OLS sense:

$$\beta_n := \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(Y_i - X_i^\top \theta\right)^2\right].$$

- From the definitions, we have for $Z_i := (\sum_i \mathbb{E}[X_i X_i^\top]/n)^{-1} X_i$:

$$\sqrt{n}\left(\hat{\beta}_n - \beta_n\right) = \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^{n} Z_i \left(Y_i - X_i^\top \beta_n\right) + o_p(1). \qquad (1)$$

- It follows: The estimate $\hat{\beta}_n$ behaves like an average. This provides a basis for asymptotically valid inference (e.g., x-y bootstrap).

- Model-Robustness: Inference justified by the linear representation (1) is valid without classical model assumptions.

Outline

Linear Regression: A harder question

Select a subset of variables $\hat{M} \subseteq \{1, 2, \ldots, p\}$ using best subset selection or lasso, say. Compute the OLS estimator $\hat{\beta}_{n,\hat{M}}$ in $\hat{M}$:

$$\begin{aligned}\hat{\beta}_{n,\hat{M}} &:= \underset{\theta \in \mathbb{R}^{|\hat{M}|}}{\arg\min} \; \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i^\top(\hat{M})\,\theta \right)^2, \\ &= \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} X_i(\hat{M}) X_i^\top(\hat{M}) \right)^{-1} \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} X_i(\hat{M}) Y_i \right)\end{aligned}$$

assuming the matrix above is invertible.

Problem

What does $\hat{\beta}_{n,\hat{M}}$ estimate? Do we need assumptions for $\hat{M}$? Do we need independent observations?

Introducing Sparsity

- Classical regression: For p $(< n)$ regressors, $\hat{\beta}_n$ satisfies

$$\left\|\hat{\beta}_n - \beta_n\right\|_2 = O_p\left(\sqrt{p/n}\right).$$

- For $p > n$ it is not possible to use all regressors due to collinearity in estimation.

- Suppose we select $\hat{M} \subseteq \{1, \ldots, p\}$ with $|\hat{M}| \leq k$ where $k < n$, but allowing $\hat{M}$ to be based on all p $(> n)$ regressors.

  $\Longrightarrow$ High-dimensional sparse regression!

- $\hat{\beta}_{n,\hat{M}}$ estimates the random target $\beta_{n,\hat{M}}$! But how?

  Significant triviality: $\left\|\hat{\beta}_{n,\hat{M}} - \beta_{n,\hat{M}}\right\|_2 \leq \sup_{|M|\leq k} \left\|\hat{\beta}_{n,M} - \beta_{n,M}\right\|_2.$

Definitions

For function $f(x, y)$, where $x \in \mathbb{R}^p$, $y \in \mathbb{R}$, define:

$$\hat{\mathbb{P}}_n[f(X, Y)] = \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i), \quad \text{and} \quad \mathbb{P}_n[f(X, Y)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(X_i, Y_i)].$$

Sample

- Gram matrix:

$$\hat{\Sigma}_n := \hat{\mathbb{P}}_n\left[XX^\top\right].$$

- "Covariance" Vector:

$$\hat{\Gamma}_n := \hat{\mathbb{P}}_n[XY].$$

- Estimator in model M:

$$\boxed{\hat{\beta}_{n,M} := (\hat{\Sigma}_n(M))^{-1}\hat{\Gamma}_n(M).}$$

Population

- Gram matrix:

$$\Sigma_n := \mathbb{P}_n\left[XX^\top\right].$$

- "Covariance" Vector:

$$\Gamma_n := \mathbb{P}_n[XY].$$

- Target in model M:

$$\boxed{\beta_{n,M} := (\Sigma_n(M))^{-1}\,\Gamma_n(M).}$$

General Result: Deterministic Inequality

Definitions:

$$\mathcal{D}_n(k) = \max_{|M|\le k} \left\|\hat{\Gamma}_n(M) - \Gamma_n(M)\right\|_2, \quad \mathrm{RIP}_n(k) = \max_{|M|\le k} \left\|\hat{\Sigma}_n(M) - \Sigma_n(M)\right\|_{op}$$

Theorem

Let $n, k \ge 1$ be integers such that $RIP_n(k) \le \lambda_{\min}(\Sigma_n)/2$. Then,

$$\sup_{|M|\le k} \left\|\hat{\beta}_{n,M} - \beta_{n,M}\right\|_2 \le C \left[\mathcal{D}_n(k) + RIP_n(k)\right],$$

for some constant C depending only on Σ_n.

Under independence or functional dependence, and sub-Gaussianity:

$$\max\left\{\mathcal{D}_n(k), \mathrm{RIP}_n(k)\right\} = O_p\left(\sqrt{\frac{k\log(ep/k)}{n}}\right).$$

Implications of General Result

- Under independence or functional dependence ($k \geq 1$):

$$\sup_{|M| \leq k} \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2 = O_p\left(\sqrt{\frac{k \log(ep/k)}{n}} \right).$$

- For a data-dependent regressor subset $\hat{M}$, the estimator $\hat{\beta}_{n,\hat{M}}$ is consistent for its random target $\beta_{n,\hat{M}}$ at the rate $\sqrt{k \log(ep/k)/n}$.

Outline

Asymptotic Uniform Linear Representation

- Similar to the uniform consistency result, under independence or functional dependence, uniformly over $|M| \le k$, we have:

$$\left\|\hat{\beta}_{n,M} - \beta_{n,M} - \frac{1}{n}\sum_{i=1}^{n} Z_{i,M}\left(Y_i - X_i^\top(M)\beta_{n,M}\right)\right\|_2 = O_p\left(\frac{k}{n}\log\left(\frac{ep}{k}\right)\right),$$

  where $Z_{i,M} = (\Sigma_n(M))^{-1} X_i(M)$.

- This implies that uniformly over all k-sparse models M,

$$\sqrt{n}\left(\hat{\beta}_{n,M} - \beta_{n,M}\right) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{i,M}\left(Y_i - X_i^\top(M)\beta_{n,M}\right).$$

- Averages all over: Leverage this for asymptotically valid FWER control over all $|M| \le k$ by stacking all averages on the right side...

Simultaneous Confidence Regions

- Define the t-statistic $t_{j,M}$ for the regressor $j \in M$:

$$t_{j,M}(\theta) := \sqrt{n}\left(\hat{\beta}_{n,M}(j) - \theta(j)\right)/\hat{\sigma}_{n,M}(j), \quad \theta \in \mathbb{R}^{|M|}.$$

- Define the statistic "max-$|t|$" $= \max_{|M|\le k} \max_{j\in M} |t_{j,M}(\beta_{n,M})|$,

  and let K_α be its upper α-quantile.

- K_α can be consistently estimated by the multiplier bootstrap.
  A similar procedure works under dependence.

- Define for any model M the confidence region

$$\hat{\mathcal{R}}_{n,M} := \left\{\theta \in \mathbb{R}^{|M|} : \max_{1\le j\le |M|} \left|t_{j,M}(\theta)\right| \le K_\alpha\right\},$$

- It follows that for any randomly selected model $\hat{M}$ with $|\hat{M}| \le k$,

$$\liminf_{n\to\infty} \mathbb{P}\left(\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}}\right) \ge 1 - \alpha.$$

Towards "Optimal" Confidence Regions

- The key component of PoSI: the max-$|t|$ statistic given by

$$\text{max-|t|} = \max_{|M| \le k} \max_{j \in M} |t_{j,M}(\beta_{n,M})|.$$

  This is the statistic used by Berk et al. (2013) (for OLS) and by Bachoc et al. (2016) (for general M-estimators).

- **Flaw:** This statistic does not account for the hierarchy of models. Smaller models should have smaller confidence regions.

- **Solution**: Treat each model size $|M|$ separately, then pool.

  Suitably modified confidence regions have size that scales "optimally" with $|M|$. (Work in progress.)

Outline

Conclusions

- We have studied linear OLS regression allowing for both misspecification and data-dependent regressor selection.
- The observations ($i = 1, \ldots, n$) are allowed to be dependent and non-identically distributed. This unification was made possible by deterministic inequalities.
- In all these settings we also provide inference tools based on high-dimensional multiplier bootstrap.
- The method of inference is computationally intensive and is provably NP-hard, but for linear regression there exist computationally efficient methods (Kuchibhotla et al. 2017).
- Finally, we note that everything mentioned here applies to a large class of M-estimators, including GLMs.

References

[1] Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2016).
Uniformly valid confidence intervals post-model-selection.
arxiv.org/abs/1611.01043.

[2] Berk, R., Brown, L. D., Buja, A., Zhang, K., Zhao, L. (2013)
Valid post-selection inference.
Ann. Statist. 41, no. 2, 802–837.

[3] Benerjee, D., Kuchibhotla A. K., and Mukherjee, S. (2018)
Large deviation and Non-uniform CLT for sums of independent high-dimensional vectors
Forthcoming to arxiv.

[4] Kuchibhotla, Brown, Buja, Berk, Zhao, George (2017)
Valid Post-selection Inference in Assumption-lean Linear Regression.
statistics.wharton.upenn.edu/research/research-listing/.

[5] Kuchibhotla, Brown, Buja, Zhao, George (2018)
A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference.
arxiv.org/abs/1802.05801.

Thank You
Questions?