Notes 2: Inference after Model Selection or Regularization

1. Introduction

Introduction

Machine learning (ML)/High-Dimensional methods provide exciting tools for dealing with big data, but

- ML methods designed with forecasting/description in mind
- Naive application may be highly misleading when inference for features of a model is the goal

For example, many economic analyses rely on observational data and aim to infer some sort of "treatment effect"

- Often want to control for other factors
  - make exogeneity of "treatment" more plausible
  - conditional object is object of interest
- Fundamental question: What controls should be used?
  - Model selection mistakes may result in invalid estimates of the effect of treatment

Problems for Inference

Two problems with employing ML methods and then doing inference for model parameters:

1. ML methods are "regularized"
   - Informative unstructured learning in highly complex models is impossible
   - Regularization refers to introducing additional structure to solve an ill-posed inverse problem
   - Fundamentally involves making a bias/variance tradeoff
     - More structure/regularization increases bias
     - More structure/regularization decreases variance
   - Regularization bias contaminates estimation of model parameters

Problems for Inference

Two problems with employing ML methods and then doing inference for model parameters:

2. ML methods are easy to overfit
   - To try to keep bias contained, might try to use mild regularization
   - Leads to larger variance in forecasts (termed overfitting)
   - With flexible/exotic ML methods hard to know you're not overfitting
   - Intuitively, overfitting means that model fits are related not just to true features of the model but also idiosyncratic unobservables $\approx$ endogeneity
   - Overfitting introduces additional bias into estimates of model parameters

2. Inference in Semiparametric Problems

Semiparametric Problems

Let's first consider a semiparametric problem:

- Prespecified low-dimensional target parameter of interest (α_0)
- To learn α_0, need to learn a high-dimensional nuisance parameter (η_0)

Canonical examples:

- Interested in specific coefficient(s) in a linear model
- Interested in an average treatment effect

Resolving Problems for Semiparametric Inference

Two useful ingredients for inference about model parameters in high-dimensions:

1. Base estimation on "orthogonal" estimating equations to alleviate regularization bias
2. Use method which provably controls overfitting or sample splitting to alleviate bias introduced by overfitting
   - Sensible ways to choose regularization in low-dimensional models and for prediction (and some high-dimensional models)

2.A. Orthogonal Estimating Equations

Orthogonal Estimating Equations

Let α_0 be the parameter of interest.

Suppose α_0 identified by moment condition

$$\mathrm{E}[\psi(W, \alpha_0, \eta_0)] = 0$$

- W denotes data
- η denotes high-dimensional nuisance parameters

Moment conditions are orthogonal if

$$\partial_\eta \mathrm{E}[\psi(W, \alpha_0, \eta)]|_{\eta=\eta_0} = 0$$

- ∂_η is an appropriate functional derivative
- Essentially means that moment condition used to learn α_0 is not violated by small perturbations of η away from η_0

Example: Partially Linear Model

Data (y_i, d_i, z_i), $i = 1, \ldots, n$, independent across i,

$$y_i = d_i\alpha_0 + g_0(z_i) + \zeta_i, \quad \mathrm{E}[\zeta_i \mid z_i, d_i] = 0$$
$$d_i = m_0(z_i) + v_i, \quad \mathrm{E}[v_i \mid z_i] = 0$$

- α_0 denotes the parameter of interest
- $g_0(z_i)$ is a nuisance parameter
- $m_0(z_i)$ is a nuisance parameter

Two potential moment conditions for learning α_0 (there are others):

1. $\mathrm{E}[\phi(W, \alpha, \eta = g)] = \mathrm{E}[(Y - D\alpha - g)D] = 0$ (non-orthogonal)
2. $\mathrm{E}[\psi(W, \alpha, \eta = (\ell, m))] = \mathrm{E}[((Y - \ell) - (D - m)\alpha)(D - m)] = 0$ (orthogonal)
   - Now have two nuisance functions: $\ell(Z) = \mathrm{E}[Y|Z]$ and $m(Z) = \mathrm{E}[D|Z]$
   - Relationship to treatment effects literature with binary D: $\ell(Z)$ is the regression function, $m(Z)$ is the propensity score

Example: Partially Linear Model

Think linear case:

$$y_i = \alpha_0 d_i + z_i' \beta_0 + \zeta_i \;\; ; \;\; d_i = z_i' \gamma_0 + v_i$$

Approach 1: When z_i is low-dimensional, OLS estimator of α_0 is numerically equivalent to OLS estimator from

$$y_i - \widehat{y}_i = \alpha_0 (d_i - \widehat{d}_i) + \zeta_i$$

- $\widehat{y}_i$ is the i^{th} element of $\widehat{Y} = P_Z Y$ and $\widehat{d}_i$ is defined similarly
- Equivalent to using moment condition 2.

Example: Partially Linear Model

Think linear case:

$$y_i = \alpha_0 d_i + z_i' \beta_0 + \zeta_i \;\; ; \;\; d_i = z_i' \gamma_0 + v_i$$

Approach 2: When z_i is low-dimensional, OLS estimator of α_0 is numerically equivalent to OLS estimator from

$$y_i - z_i' \widehat{\beta} = \alpha_0 d_i + \zeta_i$$

- $\widehat{\beta}$ is the OLS estimator of β_0
- Equivalent to using moment condition 1.

When you regularize, Approach 1 and Approach 2 are not generally equivalent.

Example: Partially Linear Model

Consider regularization via variable selection

Condition 1 corresponds (heuristically) to

1. Remove correlation between Y and D and X and D - call the new variables $\tilde{Y}$ and $\tilde{X}$.
2. Select variables from $\tilde{X}$ by finding any elements of $\tilde{X}$ that predict $\tilde{Y}$
3. Estimate α_0 by regressing Y on D and any selected elements of X

- Problem: Association between X and D leads to potential omitted variables bias.

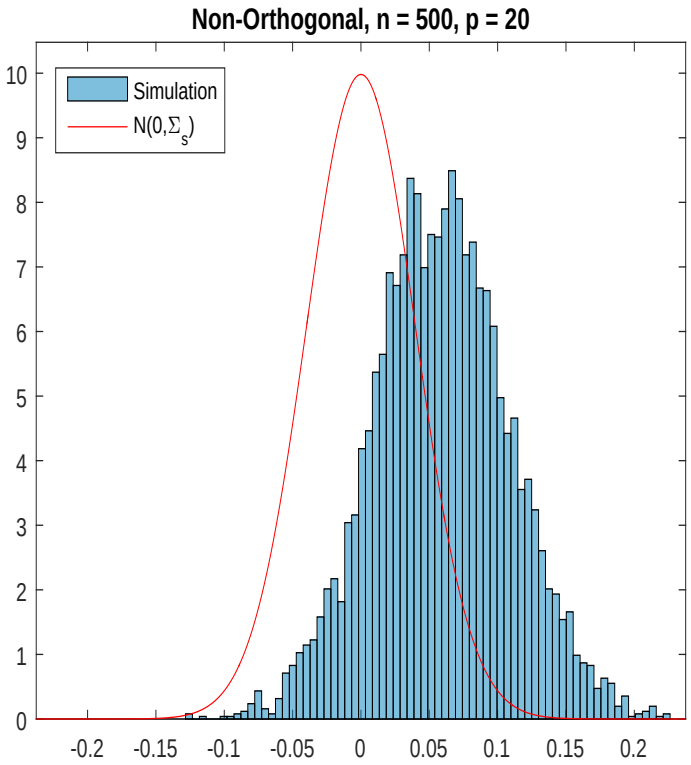
Condition 2 corresponds to

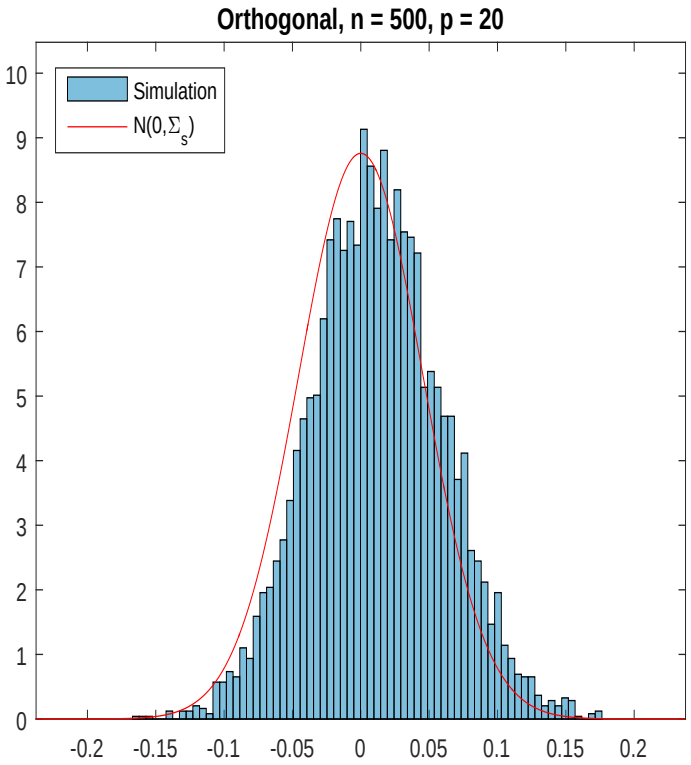
1. Select variables from X that predict Y
2. Select variables from X that predict D
3. Estimate α_0 by regressing Y on D and any variables that predict Y or D

- Alleviates potential for omitted variables bias by explicitly addressing correlation between D and X

Simulation Evidence

Results from simulation using moment conditions 1. and 2. (with nuisance parameters estimated using random forests)





2.B. Bias from Overfitting

Problem with Overfitting

Overfitting in estimating nuisance functions also leads to strong bias in estimators of parameters of interest.

Illustrate this bias again in partially linear model using a contrived example.

Example:

- Suppose $\widehat{\ell}(z_i) = \ell_0(z_i) + (y_i - \ell_0(z_i))/n^{1/2-\epsilon}$
  - $(y_i - \ell_0(z_i))/n^{1/2-\epsilon}$ captures overfitting - estimator is truth plus a term that depends on the idiosyncratic realizations in our observed sample
  - Estimator $\widehat{\ell}(z_i)$ is excellent in terms of being consistent and converging at essentially the $\sqrt{n}$ rate (which is infeasible in high-dimensional/nonparametric settings)

Bias from Overfitting

What happens to the estimator of α_0 based on the orthogonal moment condition now?

- Assume we know $m_0(\cdot)$.
- Recall that $d_i - m_0(z_i) = v_i$.
- Recall that $y_i = \alpha_0 d_i + g_0(z_i) + \zeta_i = \ell_0(z_i) + \alpha_0 v_i + \zeta_i$ where $\ell_0(z_i) = \alpha_0 m_0(z_i) + g_0(z_i)$
  - So $y_i - \widehat{\ell}(z_i) = \alpha_0 v_i + \zeta_i - (\alpha_0 v_i + \zeta_i)/n^{1/2-\epsilon}$

Substituting in to the solution $\widehat{\alpha}$ to condition 2 gives

$$\begin{aligned}\sqrt{n}(\widehat{\alpha} - \alpha_0) &= \frac{\frac{1}{\sqrt{n}}\sum_i v_i(\zeta_i - (\alpha_0 v_i + \zeta_i)/n^{1/2-\epsilon})}{\frac{1}{n}\sum_i v_i^2} \\ &= O_p(1)\left(\frac{1}{\sqrt{n}}\sum_i v_i\zeta_i - \frac{n^\epsilon}{n}\sum_i \alpha_0 v_i^2 - \frac{n^\epsilon}{n}\sum_i v_i\zeta_i\right) \\ &= O_p(1)(O_p(1) + n^\epsilon O_p(1) + o_p(1))\end{aligned}$$

which diverges (akin to endogeneity bias).

Solutions to Bias due to Overfitting

1. Use procedures that provably do not overfit
   - Traditional semiparametric approaches: Levit (1975), Ibragimov and Hasminskii (1981), Bickel (1982), Robinson (1988), Newey et al. (1998), Newey (1990), van der Vaart (1991), Andrews (1994a), Newey (1994), Newey et al. (2004), Robins and Rotnitzky (1995), Linton (1996), Bickel et al. (1998), Chen et al. (2003), van der Laan and Rose (2011), and Ai and Chen (2012)
   - ML approaches that provably control overfitting (mostly lasso-based): Belloni, Chen, Chernozhukov, and Hansen (2012), Belloni, Chernozhukov, and Hansen (2014); Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017); Javanmard and Montanari (2014); van de Geer, Buhlmann, Ritov, and Dezeure (2014); Zhang and Zhang (2014)
   - Seem to be somewhat special and require leveraging special structure on the problem

Solutions to Bias due to Overfitting

2. Sample split

Basic idea in partially linear model.

- Suppose sample split into two parts (A and B, equal sized for simplicity)
- Part A used to estimate nuisance functions. Part B to estimate the parameter of interest
  - Model overfitting as $\widehat{\ell}^A(z_i) = \ell_0(z_i) + R^A(z_i)/(n/2)^{1/2-\epsilon}$. Note that $R^A(z_i)$ independent of everything in sample B (under independence)

What happens to the estimator $\widehat{\alpha}$ under condition 2 gives

$$\sqrt{n/2}(\widehat{\alpha} - \alpha_0) = \frac{\frac{1}{\sqrt{n/2}} \sum_{i \in B} v_i(\zeta_i - R^A(z_i)/(n/2)^{1/2-\epsilon}}{\frac{1}{n/2} \sum_{i \in B} v_i^2}$$

$$= O_p(1)\left(\frac{1}{\sqrt{n/2}} \sum_{i \in B} v_i \zeta_i - \frac{(n/2)^{\epsilon}}{n/2} \sum_{i \in B} v_i R^A(z_i)\right)$$

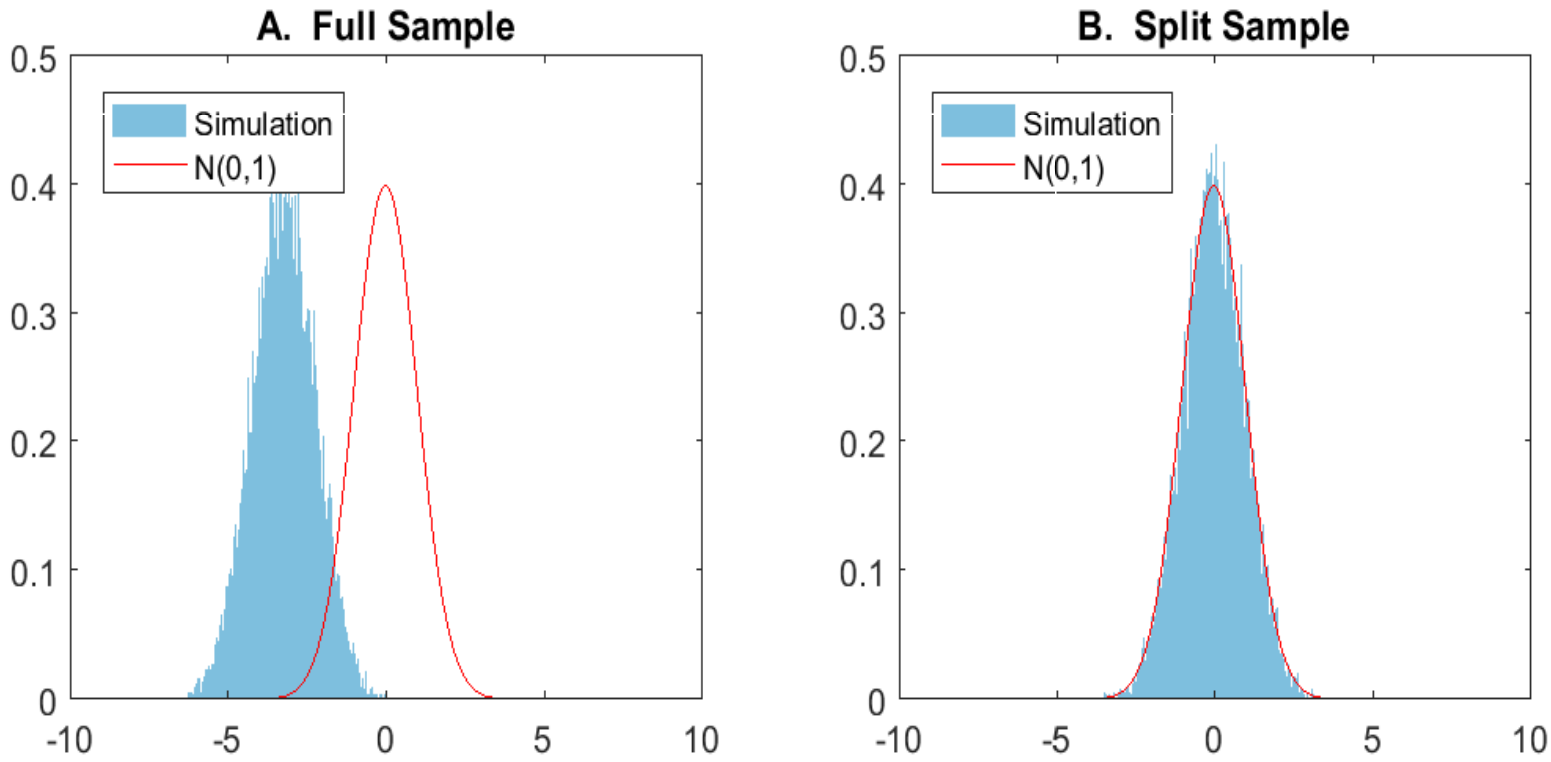
$$= O_p(1)(N(0, V) + (n/2)^{-1/2+\epsilon} O_p(1)) = O_p(1)N(0, V) + o_p(1)$$

because v_i is mean zero and independent of $R^A(z_i)$

Full efficiency can be restored by flipping roll of sample A and B and averaging the results.

Simulation Evidence

Results from simulation using moment condition 2 with and without sample splitting (with nuisance parameters estimated using random forests)



[Note: Previous figure for comparison of condition 1 and 2 used sample splitting for both estimators]

Cross-Fitting

Lots of ways to use sample-splitting to control bias from overfitting

Cross-fitting:

- Split the data into K approximately equal-sized parts
- For $k = 1, ...,$K
  - (a) Remove subsample k
  - (b) Fit high-dimensional component using remaining $K - 1$ subsamples
  - (c) Use only observations from subsample k to estimate parameter of interest using estimator of nuisance function from (b) $\rightarrow \tilde{\alpha}_k$ and estimated standard error $\tilde{\sigma}_k$
- Estimate α_0 as $\check{\alpha} = \frac{1}{K}\sum_{k=1}^{K}\tilde{\alpha}_k$ with standard error $\check{\sigma} = \sqrt{\frac{1}{K^2}\sum_{k=1}^{K}\tilde{\sigma}_k^2}$

The specific sample split also introduces variability. We can account for some of this by considering many different sample splits.

- Consider $b = 1, ..., B$ splits $\rightarrow \{\check{\alpha}_b, \check{\sigma}_b\}_{b=1}^{B}$
- Report summaries of $\{\check{\alpha}_b, \check{\sigma}_b\}_{b=1}^{B}$. E.g.
  - $\bar{\alpha} = \frac{1}{B}\sum_{b=1}^{B}\check{\alpha}_b$
  - $\sqrt{\frac{1}{B}\sum_{b=1}^{B}(\check{\sigma}_b^2 + (\check{\alpha}_b - \bar{\alpha})^2)}$

3. Semiparametric Examples

Application: Effect of Abortion on Crime Rates

Estimate the consequences of abortion rates on crime, Donohue and Levitt (2001)

$$y_{it} = \gamma_t + d_{it}\alpha_0 + x_{it}'\beta_g + \zeta_{it}$$

- y_{it} = change in crime-rate (violent, property, or murder per 1000) in state i between t and $t-1$,
- d_{it} = change in the "effective" abortion rate,
- x_{it} = controls for time-varying confounding state-level factors, including initial conditions, interactions, squared terms, and interactions of all these variables with trend and trend-squared
- γ_t time effects (not-selected over)
- $p = 284$, $n = 576$

Application: Effect of Abortion on Crime Rates

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
DL Table 4	-.129	.024	-.091	.018	-.121	.047
First-Diff	-.152	.034	-.108	.022	-.204	.068
All Controls	.006	.483	-.154	.163	2.240	2.184
Non-Orthogonal	-.155	.033	-.101	.022	-.021	.051
Orthogonal	-.089	.053	-.020	.051	-.045	.069
Orthogonal - Sample Split	-.048	.077	-.040	.042	-.075	.103

- Use lasso for fitting nuisance functions
  - Without sample-splitting, use tuning choices that theoretically provide good fit while avoiding overfitting (specifically, Belloni, Chernozhukov, Hansen, and Kozbur (2016) which accommodates clustering)
  - With sample-splitting, use iid 10-fold cross-validation
- Results in-line with the heuristic critique raised by Foote and Goetz (2008).

Application: Effect of 401(k) Participation on Accumulated Assets

Follow Poterba et al (97), Abadie (03). Data from 1991 SIPP, $n = 9,915$

- Y is net total financial assets or total wealth
- D is indicator for working at a firm that offers a 401(k) plan
- X includes age, income, family size, education, and indicators for married, two-earner, defined benefit pension, IRA participation, and home ownership

D is perhaps plausibly exogenous at the time when 401(k) was introduced

Important to control for income to capture unobserved heterogeneity in employment decisions (Poterba, Venti, and Wise 94, 95, 96, 01)

- How do we specify what it means to control? Is our functional form right?

ATE with Heterogeneous Treatment Effects

Estimate of ATE allowing for full heterogeneity of treatment effects:

Model:

$$y_i = d_i g_1(x_i) + (1 - d_i) g_0(x_i) + \zeta_i$$
$$d_i = m(x_i) + u_i$$

Orthogonal estimating equation for ATE (α) from Hahn (1998):

$$\mathrm{E}\left[\frac{d_i(y_i - g_1(x_i))}{m(x_i)} - \frac{(1 - d_i)(y_i - g_0(x_i))}{1 - m(x_i)} + g_1(x_i) - g_0(x_i) - \alpha\right] = 0$$

Estimate nuisance functions using $g_0(\cdot)$, $g_1(\cdot)$, and $m(\cdot)$ ML methods

Application: Effect of 401(k) Participation on Accumulated Assets

Estimated Effect of 401(k) Eligibility on Net Financial Assets

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Regression Model							
ATE (2 fold)	6830	7713	7770	7806	7764	7702	7546
	[1282]	[1208]	[1276]	[1159]	[1328]	[1149]	[1360]
	(1530)	(1271)	(1363)	(1202)	(1468)	(1170)	(1533)
ATE (5 fold)	7170	7993	8105	7713	7788	7839	7753
	[1201]	[1198]	[1242]	[1155]	[1238]	[1134]	[1237]
	(1398)	(1236)	(1299)	(1177)	(1293)	(1148)	(1294)
B. Partially Linear Regression Model							
ATE (2 fold)	7717	8709	9116	8759	8950	9010	9125
	[1346]	[1363]	[1302]	[1339]	[1335]	[1309]	[1304]
	(1749)	(1427)	(1377)	(1382)	(1408)	(1344)	(1357)
ATE (5 fold)	8187	8871	9247	9110	9038	9166	9215
	[1298]	[1358]	[1295]	[1314]	[1322]	[1299]	[1294]
	(1558)	(1418)	(1328)	(1328)	(1355)	(1310)	(1312)

4. Conditional Inference

Conditional Inference

Not all problems fall within the semiparametric framework

Some good examples come in the context of policy evaluation with heterogeneous conditional average treatment effects (CATE(x) = $\mathrm{E}[Y_1 - Y_0 | X = x]$)

- E.g. Suppose $D(x)$ is a new policy based on what you've learned such as $D(x) = 1(CATE(x) > 0)$. Want to evaluate $\mathrm{E}_{D(x)}[Y]$ (mean of Y under counterfactual that $D(x)$ is adopted)
- E.g. Might want to test CATE(x^*) > 0 for some x^* chosen after looking at the data

Rely on using the data to learn what you want to test - Object of interest not prespecified before seeing the data

Conditional Inference

Inference problem is easy if we are willing to

1. Split the sample
2. Learn the object of interest using only sample A
3. Condition on the answer from sample A
4. Do inference using only sample B

Works because object of interest is prespecified from standpoint of sample B - standard inference problem

Drawbacks

- have to "commit" to answer from sample A
- only use a subset of the data - higher variance

5. Conditional Inference Example

JTPA Data

JTPA (Job Training Partnership Act) Experiment:

- males randomly assigned offer of JTPA training services ($n = 5102$)
- outcome: total earnings over the 30 month period following treatment assignment (average $19,147)
- controls
  - high school dummy
  - black and hispanic dummies
  - worked at least 12 weeks in 12 months prior to assignment dummy
  - 5 age dummies
  - earnings from second follow-up survey dummy

Split sample into two equal halves - estimating and testing sample

Evaluating Treatment Policies

Consider simple set-up where we want to do inference on which to implement.

Use several different ML methods to estimate CATE(x).

- lasso (with fully saturated model)
- tree
- random forest
- boosted trees

Several different policies:

- Treat no one
- Treat everyone
- Treat if CATE(x) > 0 for each CATE estimator
- Treat if CATE(x) > 0 for majority of CATE estimators
- Treat if average CATE(x) across estimators is positive

Use the bootstrap to do inference using data from sample B

Results

Estimated Average outcomes under each policy

Policy	Average Outcome
None	14991.34
All	16085.51
Lasso - CV	16081.01
Lasso - Plug-in	16090.79
Tree	16088.16
Forest	16087.83
Boosting	16089.12
Forward Selection	16091.39
Majority	16086.94
Average	16089.17

Everything but treating no one looks pretty similar.

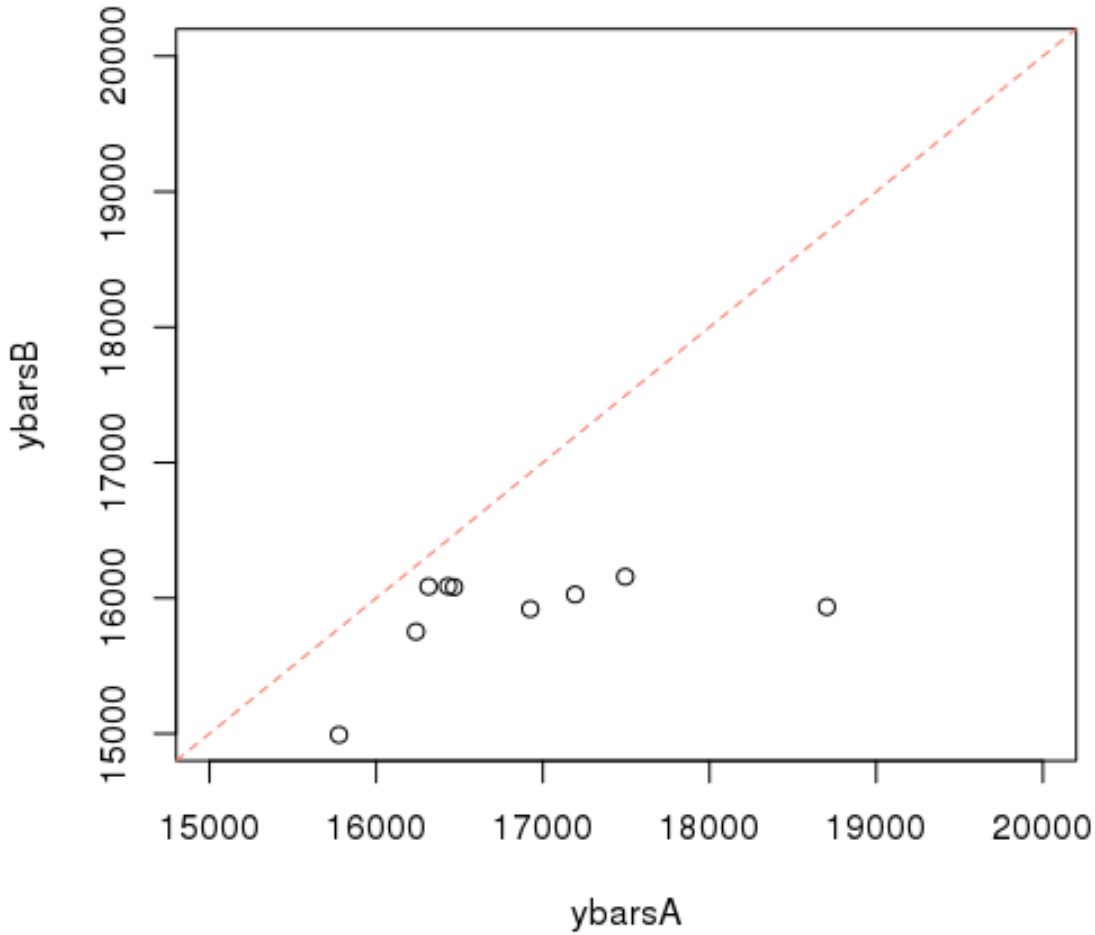
Results

Confidence intervals for difference from no treatment (multiplicity adjusted):

Policy	LB	UB
all	86.81	2101.54
lasso - cv	338.24	1850.74
lasso - plug-in	346.68	1852.20
tree	340.71	1852.93
forest	340.45	1852.52
boost	341.87	1853.68
step	344.16	1855.93
majority	339.71	1851.48
mean	341.79	1853.87

Not much evidence that worrying about heterogeneity is worth it in this example.

In-Sample vs. Out-Of-Sample Performance



Sample A is estimation sample. Sample B is testing sample.

Results

For fun, let's suppose there's a cost of $500 to offering the policy and we only want to treat people where the benefit exceeds the cost. Estimated Average outcomes under each policy

Policy	Average Outcome
None	14991.34
All	15585.51
Lasso - CV	15644.65
Lasso - Plug-in	15835.97
Tree	15710.98
Forest	15781.06
Boosting	15651.06
Forward Selection	15838.53
Majority	15670.21
Average	15707.97

Results

Confidence intervals for difference from no treatment (multiplicity adjusted):

Policy	LB	UB
all	-406.11	1594.45
lasso - cv	-92.51	1399.13
lasso - plug-in	101.77	1587.47
tree	-26.33	1465.59
forest	43.81	1535.63
boost	-85.75	1405.20
step	101.69	1592.69
majority	-66.65	1424.39
mean	-29.01	1462.26

Maybe a little more going on here

Results

Let's look at fraction treated under each rule:

Policy	Average Outcome
None	0.00
All	1.00
Lasso - CV	0.88
Lasso - Plug-in	0.51
Tree	0.75
Forest	0.61
Boosting	0.88
Forward Selection	0.51
Majority	0.83
Average	0.76

6. "Nonparametric" Inference

"Nonparametric" Inference

Not all problems are readily treated in the two previous paradigms

- Might be truly nonparametric
- Do not want conditional inference
- Finding orthogonal moment functions may be difficult

Examples:

- $\mathrm{E}[Y_1 - Y_0|X = x]$ =CATE(x)
- Expected profit from implementing optimal strategy when cost of treatment is c =
  $\mathrm{E}[Y_0|D^*(X) = 0]\Pr(D^*(X) = 0) + \mathrm{E}[Y_1 - c|D^*(X) = 1]\Pr(D^*(X) = 1)$
  where $D^*(X) = 1(\text{CATE}(X) > c)$

Sparsity

With $p \approx n$ or $p \gg n$, cannot do inference for many objects without further assumptions

For this section, we will stop being agnostic about structure and impose (approximate) **sparsity**:

- Let θ ($p \times 1$) be the (full) model parameter
- Assume $\|\theta_0\|_0 = s_n \ll n$
- Estimation problem becomes find and estimate non-zero elements of θ_0
  - lasso/post-lasso, forward selection, ...
- Do not impose $\beta_{\min}$-like conditions
  - include models where perfect model selection impossible

Simple Example

Consider the model

$$\boxed{y_i = w_i'\beta_0 + T_i(w_i'\gamma_0) + \epsilon_i = x_i'\theta_0 + \epsilon_i}$$

with $\theta_0 = (\beta_0', \gamma_0')'$, $\|\theta_0\|_0 \leq s$, $\text{supp}(\theta_0) = S$, $\mathrm{E}[\epsilon_i | x_i] = 0$.

Consider an inference target:

$$a(\theta_0)$$

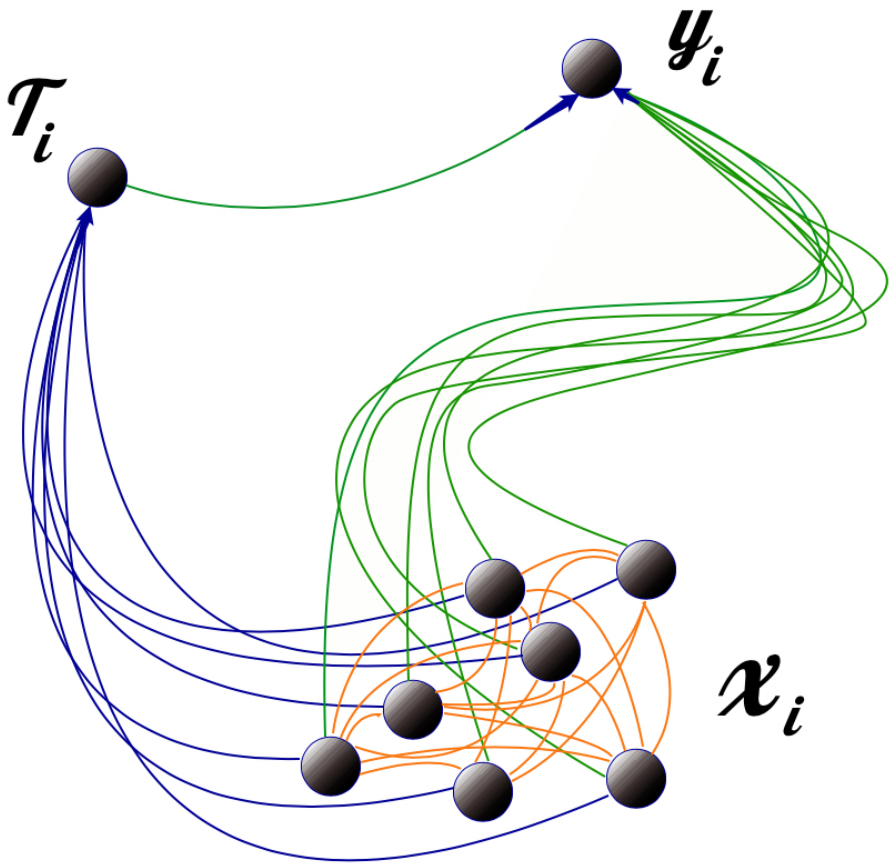
for some fixed $a \in \{\mathbb{R}^p \to \mathbb{R}\}$.

E.g.

- $a(\beta) = x_0'\beta$ for some x_0
- $a(\beta) = x_0'\gamma$ for some x_0
- $a(\beta) = \sum_{j=1}^p \beta_j$
- $a(\beta) = \max_j |\gamma_j|$

Note that $a(\cdot)$ may depend on x_i or other objects and need not be linear.

What We Need to Learn



All arrows relevant: no parameters are nuisance parameters

Potential Procedure

One way to construct a confidence set $\mathcal{I}$ for $a(\theta_0)$ if we knew an upper-bound $\overline{s}$ on s:

- For each $K \subset \{1, ..., p\}$, where $|K| \leq \overline{s}$, form asymptotic $1 - \alpha$ confidence interval $[\ell_K, u_K]$ supposing that $K = \text{supp}(\theta_0) \equiv S$ is correctly specified
- Set $\mathcal{I} = \bigcup_K [\ell_K, u_K]$
- This yields correct coverage

$$\mathrm{P}\{a(\theta_0) \in \mathcal{I}\} \geq \mathrm{P}\{a(\theta_0) \in [\ell_S, u_S]\} \geq 1 - \alpha + o(1)$$

Potential Procedure

Alternatively, we can perform the following optimizations and define $\mathcal{I} = [\ell, u]$ with

- $\ell = \min_{b \in \mathbb{R}^p} a(b) - c_{1-\alpha}\textbf{s.e.}(a(b))$

$$s.t.\ \|b\|_0 \leq \overline{s}, \quad b = \textbf{reg.coeff}(y_i \text{ on } x_{i,\text{supp}(b)})$$

- $u = \max_{b \in \mathbb{R}^p} a(b) + c_{1-\alpha}\textbf{s.e.}(a(b))$

$$s.t.\ \|b\|_0 \leq \overline{s}, \quad b = \textbf{reg.coeff}(y_i \text{ on } x_{i,\text{supp}(b)})$$

[For instance, using delta method for **s.e.**$(a(b))$]

Potential Procedure

This approach yields correct coverage, but has at least 2 major drawbacks:

- Computationally difficult for even moderate $\overline{s}$
- Gives VERY large confidence regions

Potential Procedure

The optimization is **not local** to the true model:

- If the set $K \subset \{1, ..., p\}$ consists of irrelevant regressors which are orthogonal to true regressors, the corresponding estimate **reg.coeff**(y_i on $x_{i,K}$) will be approximately 0 with high probability
- Under this case, the interval $[\ell_K, u_K] \ni 0$ with high probability
- The resulting inference has no power against some fixed alternatives

Proposed Procedure

"Local" procedure:

Step 1.
Use Lasso (or other) to perform model selection y_i on x_i:

$$\rightarrow \widehat{S}^{(0)}$$

Proposed Procedure

Step 2.

- Define $\widehat{S}^{(\text{low})}$ by the following program:

$$\widehat{S}^{(\text{low})} = \arg \min_{K \subseteq \{1,\ldots,p\}} a(b) - c_{1-\alpha}\textbf{s.e.}(a(b))$$

$$\text{s.t. } |K| \leq \overline{s},\ K \supseteq \widehat{S}^{(0)},$$
$$b = \textbf{reg.coeff}(y_i \text{ on } x_{i,K}).$$

- Define $\widehat{S}^{(\text{up})}$ similarly.

Proposed Procedure

Two drawbacks:

- Cannot rule out $\exists j \notin S : j$ selected. Then *all* subsets $\widehat{S}^{(0)} \subseteq K \subseteq \{1, ..., p\}$ contain some $j \notin S$. This means we cannot leverage a central limit theorem for $\sum_{i=1}^{n} x'_{i,S}\varepsilon_i$ since S was not even considered in the optimization. Furthermore,

$$\mathrm{E}[x'_{ij}\varepsilon_i | j \text{ selected}, j \notin S] \neq 0.$$

- Though much faster than the original (non-local) optimization, it can still be computationally infeasible.

Proposed Procedure

To address the first problem, we require $\overline{s} \geq |T|$ where $T \subseteq \{1, ..., p\}$ such that

$$\mathrm{P}(T^c \cap \widehat{S}^{(0)} \neq \emptyset) = o(1)$$

- Essentially states that set of variables liable to be falsely selected into initial model is not too large
- Can be shown to hold under design conditions

-OR-

Sample split:

- Partition $\{1, ..., n\}$ into $A \sqcup B$. Estimate $\widehat{S}^{(0)}$ on sample A. Construct union of confidence intervals on sample B.
- Allows Gaussian inference

Formal conditions under which this all works exist.

- Essentially, need that asymptotic inference if you knew the true model would work
- Mild regularity on functional $a(\cdot)$ - complexity of $a(\cdot)$ shows up in rates of convergence

Proposed Procedure

To address computational problem

- Solve the two optimizations approximately
- Use greedy algorithm
  - e.g. forward selection (we do this in the simulations and empirical example)
- Other approaches are also possible, eg. semidefinite relaxations

Preliminary Simulation Evidence

Table: Inference Simulation [$n = 1000,\ p = 501$] Results: β_1

	Coverage Prob.	Interval Length
True Support	0.95	0.83
All	0.94	1.17
Oracle-Style	0.00	0.01
Lasso + $\log(n)$	1.00	1.06
Lasso + $n^{1/3}$	1.00	1.14
Lasso + $n^{1/2}$	1.00	1.54
Projection	0.95	0.83

Preliminary Simulation Evidence

Table: Inference Simulation [$n = 1000,\ p = 501$] Results: $g(x_0)$

	Coverage Prob.	Interval Length
True Support	0.95	0.46
All	0.95	2.92
Oracle-Style	0.00	0.21
Lasso + $\log(n)$	1.00	2.28
Lasso + $n^{1/3}$	1.00	2.89
Lasso + $n^{1/2}$	1.00	5.76
Projection	1.00	238.40

Preliminary Simulation Evidence

Table: Inference Simulation [$n = 1000,\ p = 501$] Results: $h(x_0)$

	Coverage Prob.	Interval Length
True Support	0.94	0.62
All	0.95	4.13
Oracle-Style	0.00	0.28
Lasso + $\log(n)$	1.00	3.26
Lasso + $n^{1/3}$	1.00	4.13
Lasso + $n^{1/2}$	1.00	8.25
Projection	1.00	283.04

Preliminary Simulation Evidence

Table: Inference Simulation [$n = 1000,\ p = 501$] Results: $\mathrm{E}[\Delta\pi(D^*)]$

	Coverage Prob.	Interval Length
True Support	0.89	0.16
All	0.00	0.17
Oracle-Style	0.00	0.14
Lasso + $\log(n)$	0.99	0.27
Lasso + $n^{1/3}$	0.99	0.29
Lasso + $n^{1/2}$	1.00	0.40
Projection	–	–

7. "Nonparametric" Examples

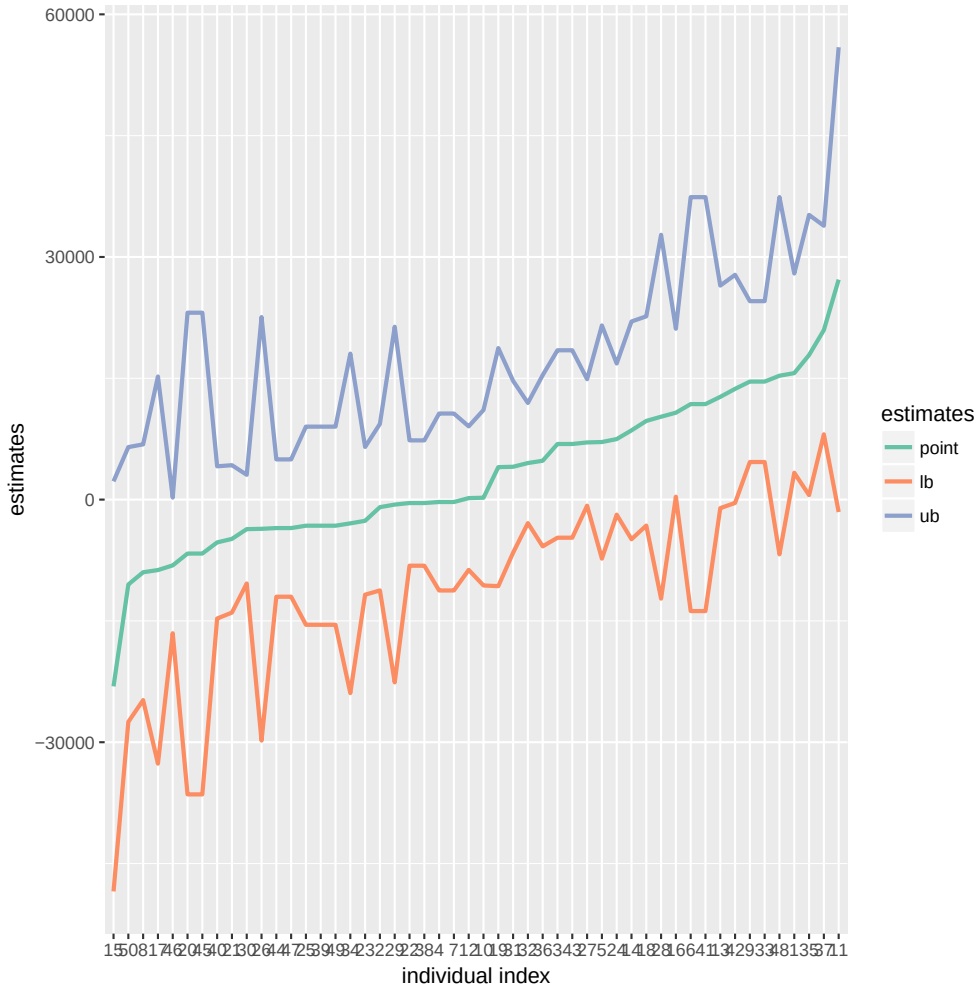
Example 1: JTPA Data

Suppose we are interested in actual individual specific treatment effects.

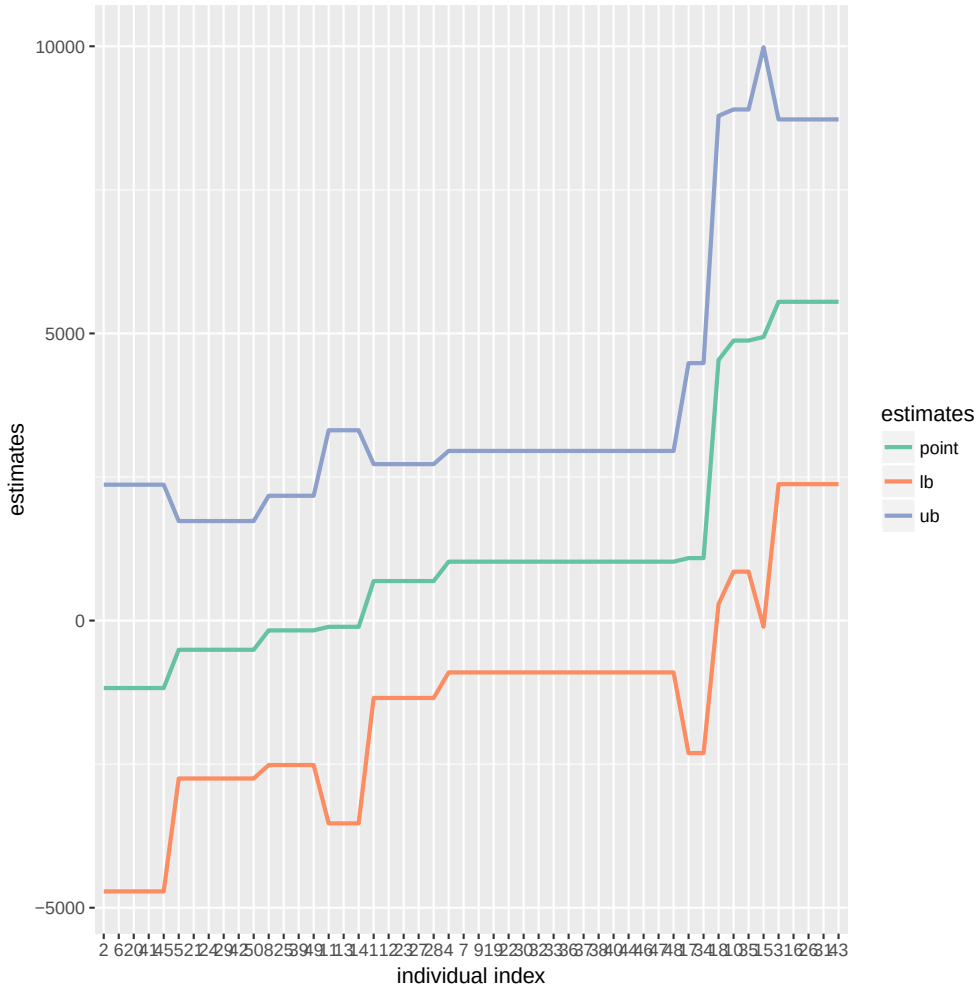
We report intervals for 50 observations from

- OLS with all interactions (with near collinear columns removed by default in R)
  - s.e.'s are potentially badly off - (jackknife fails - returns "not a number")
- Selection based estimates assuming perfect selection
- Forward selection procedure

OLS Results

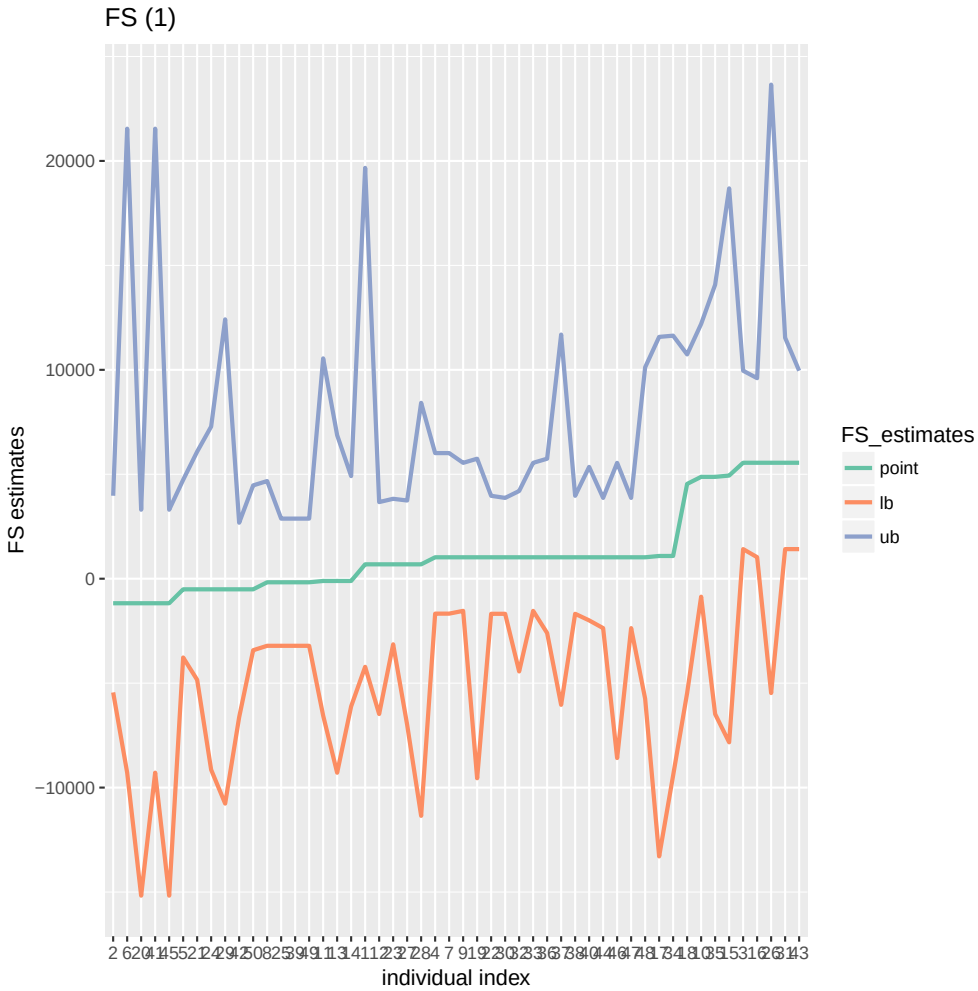


Oracle Results



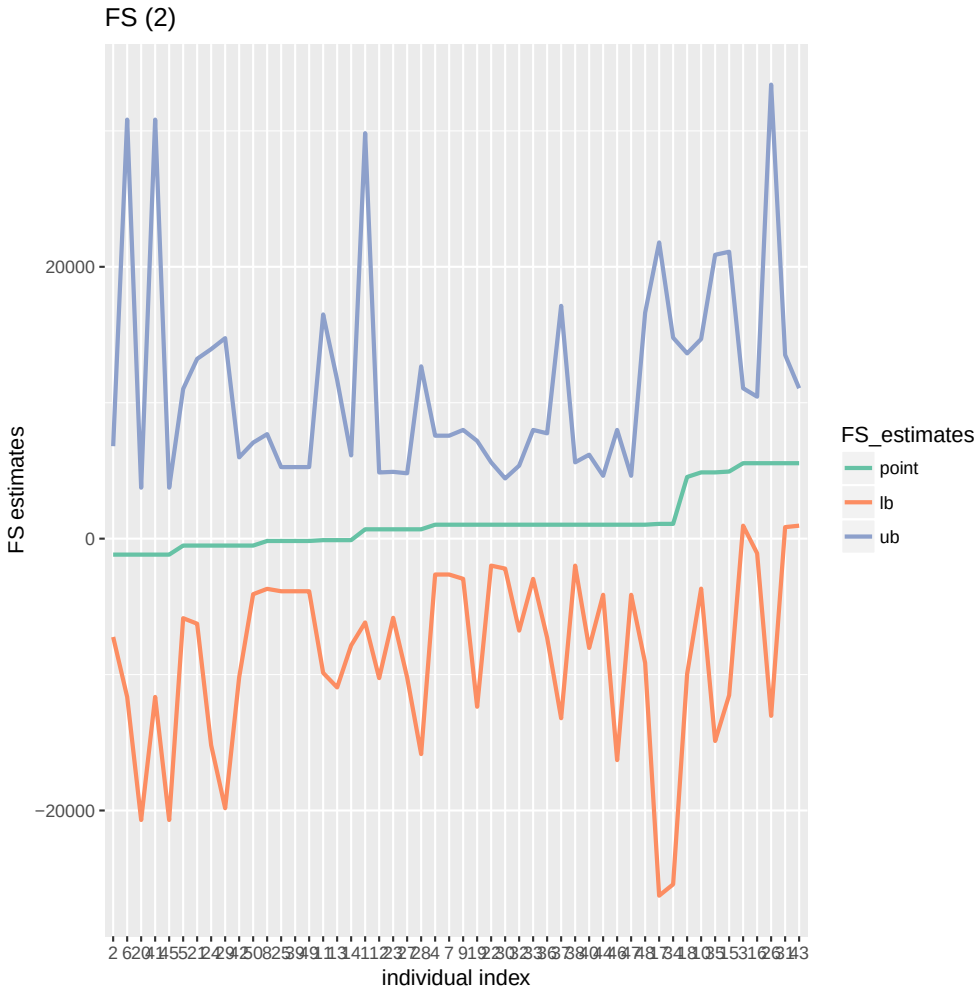
Undersmoothing

One additional variable:



Undersmoothing

2 additional variables:

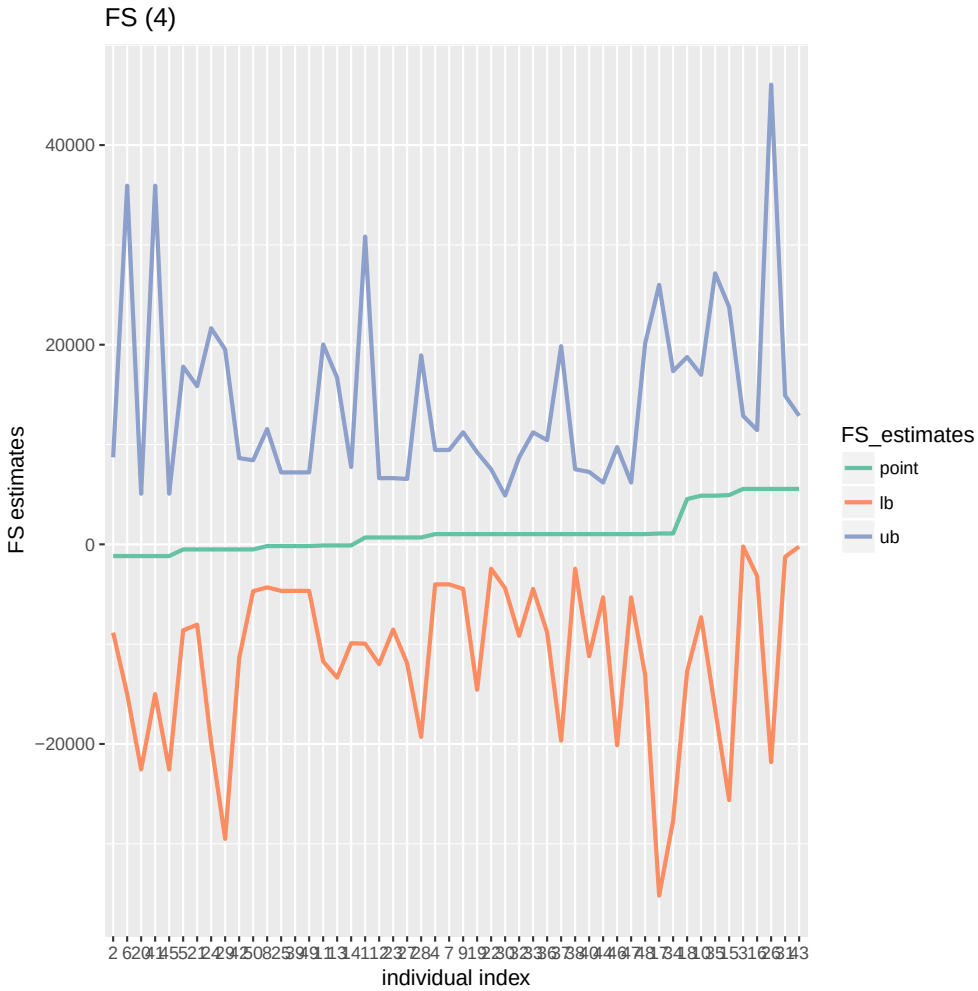


Undersmoothing

3 additional variables:

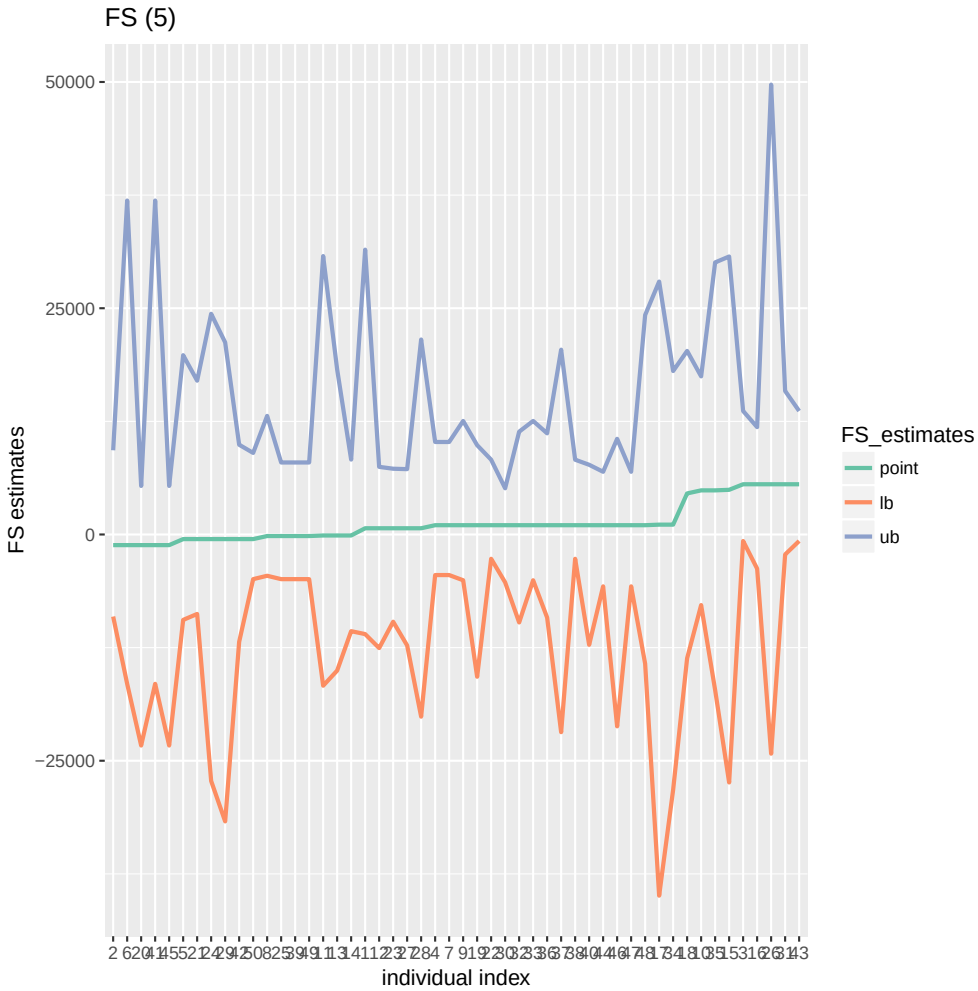
Undersmoothing

4 additional variables:



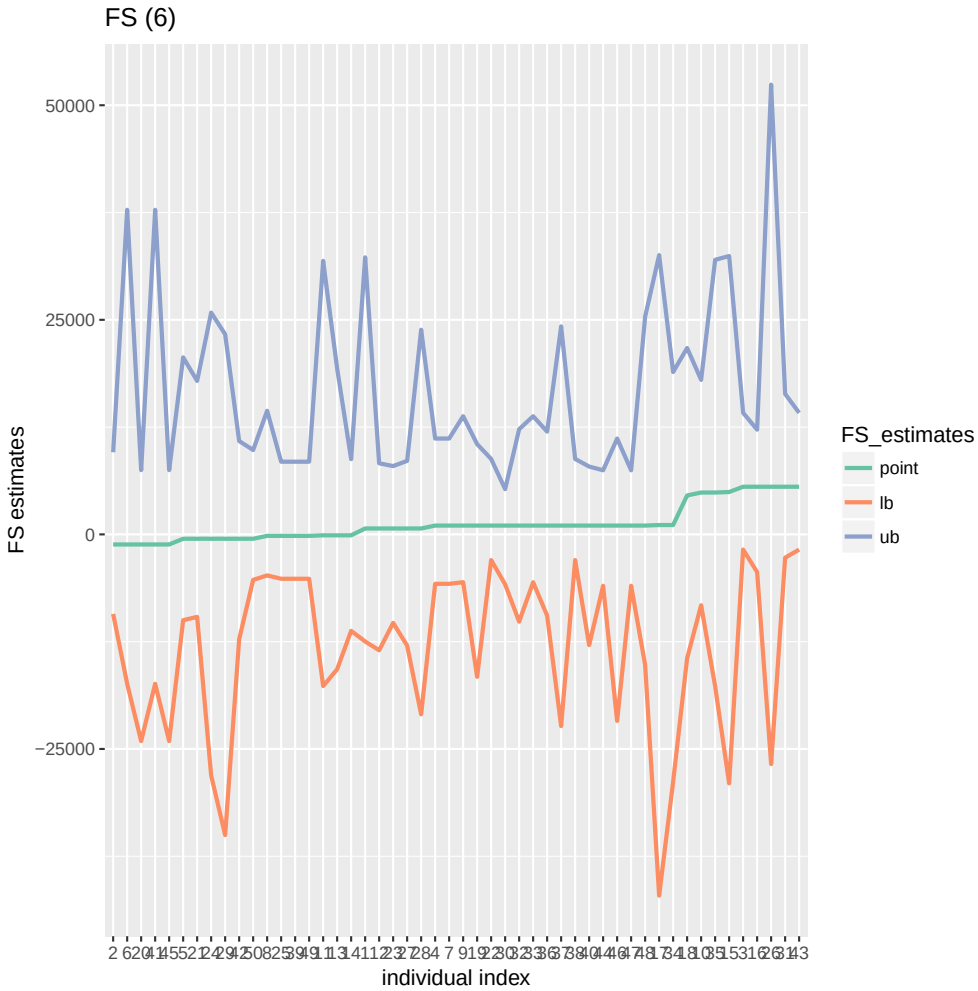
Undersmoothing

5 additional variables:



Undersmoothing

6 additional variables:

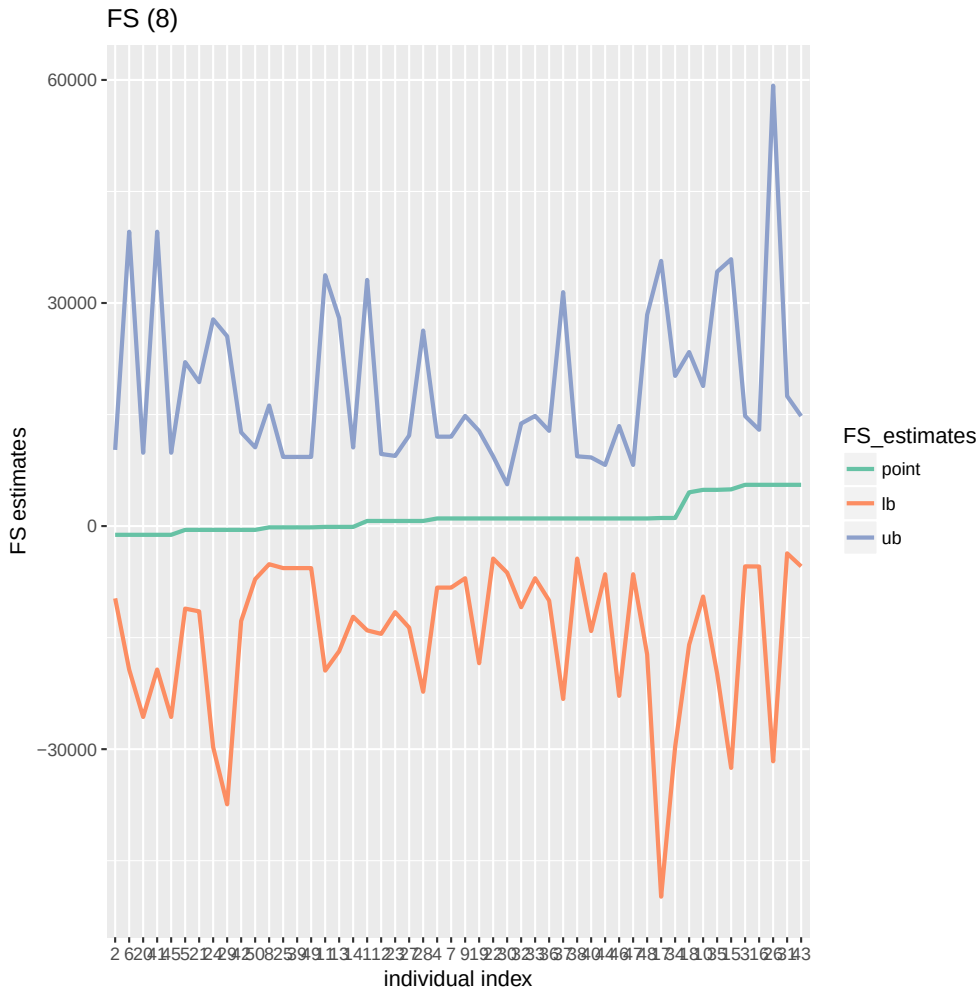


Undersmoothing

7 additional variables:

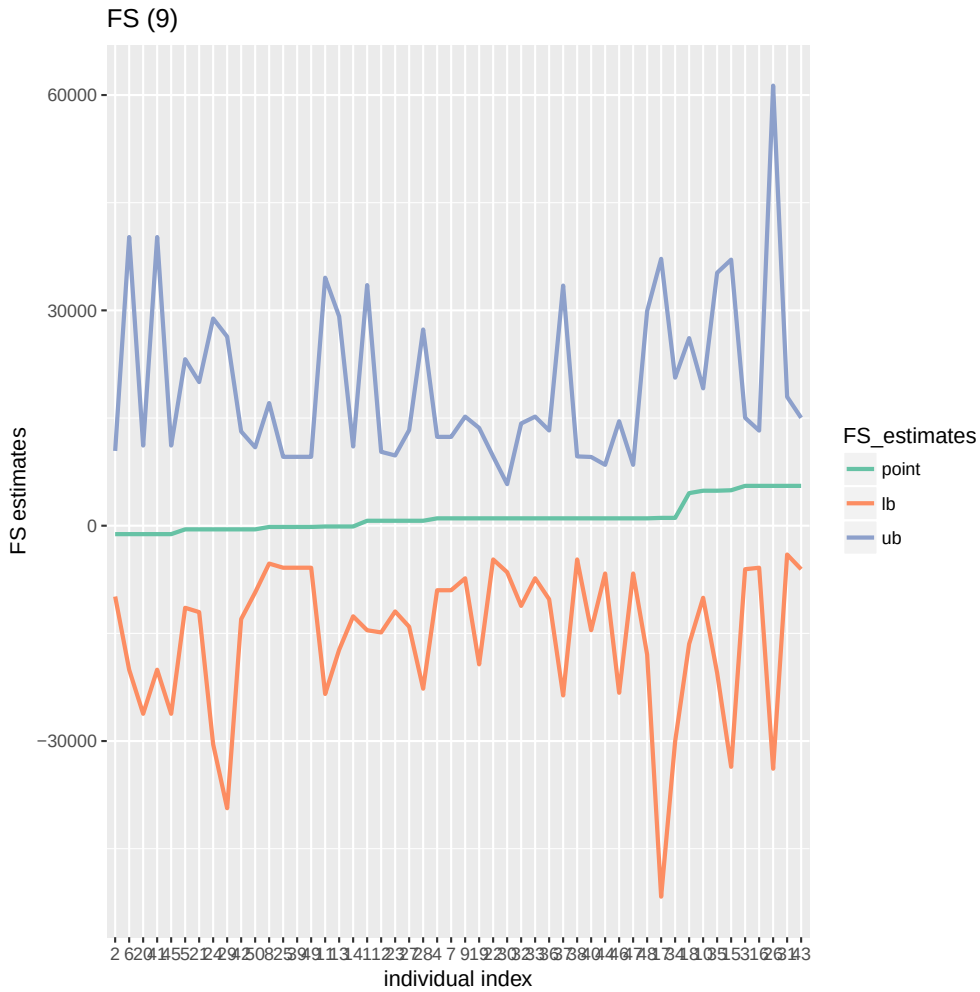
Undersmoothing

8 additional variables:



Undersmoothing

9 additional variables:



Example: Targeted Marketing

Data:

- y_i: Total expenditure
- x_i: Consumer i's characteristics
  - Demographics
  - Purchasing history
- d_i: Whether consumer i received a promotion

We have $n = 287970$, $p = 2139$ (Characteristics and interactions with treatment indicator). d_i is assigned randomly with treatment probability of 2/3.

Model

Model:

$$y_i = x_i' \beta_0 + (d_i \cdot x_i)' \gamma_0 + \varepsilon_i$$

- Cost of promotion:

$$C_i = \bar{C} + \eta_i$$
$$\eta_i \sim N(0, \sigma^2)$$

- The firm gets fixed margin M from sales.

Profit

Expected profit from strategy d:

$$\begin{aligned}\mathrm{E}[\pi(d)] &= \mathrm{E}[M(\beta_0 + x'\beta + d(x,c)(\gamma_0 + x'\gamma)) - d(x,c)c] \\ &= \mathrm{E}[M(\beta_0 + x'\beta) + d(x,c)(M(\gamma_0 + x'\gamma) - c)]\end{aligned}$$

Strategy: Given known costs, strategy is $D^* = 1(M(\gamma_0 + x'\gamma) > C)$. We then integrate over distribution of C.

Profit Differential

Profit differential relative to no mailing is reported:

$$\begin{aligned} \mathrm{E}[\Delta\pi(D^*)] = \mathrm{E}[&\Phi((M(\gamma_0 + x'\gamma) - \bar{C})/\sigma) \\ &\times (M(\gamma_0 + x'\gamma) - \bar{C})] \\ &- \mathrm{E}[1(M(\gamma_0 + x'\gamma) - \bar{C} > \eta)\eta] \end{aligned}$$

- In data, we have $M \approx .3$, $\bar{C} \approx .7$, $\sigma \approx .1$.

Results

Table: Average Profit Differential $\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}\left[\pi_i^*\right]-\mathrm{E}\left[\pi_i^0\right]\right)\right)$

Estimator	Estimate	S.E.	Lower	Upper
OLS	1.4295	0.0652	1.3017	1.5573
Post-Lasso	0.3332	0.0412	0.2524	0.4140
FS(1)			0.2225	0.4524
FS(2)			0.2169	0.4739
FS(3)			0.2125	0.4866
FS(4)			0.2080	0.4993
FS(5)			0.2044	0.5100
FS(6)			0.2019	0.5231
FS(7)			0.1997	0.5318
FS(8)			0.1979	0.5399
FS(9)			0.1967	0.5463
FS(10)			0.1947	0.5536
FS(11)			0.1936	0.5598
FS(12)			0.1926	0.5654
FS(13) = FS(log(n))			0.1918	0.5717

8. Summary

Summary

Inference in high-dimensional settings complicated by

- regularization bias
- overfitting

Outlined three different paradigms for doing inference for different objects that address these complications.