Regularized Treatment Effect Estimation

Carlos M. Carvalho (UT Austin)
P. Richard Hahn (Chicago Booth)
David Puelz (UT Austin)

Brown Bag, May 2016

Moving past 1930's

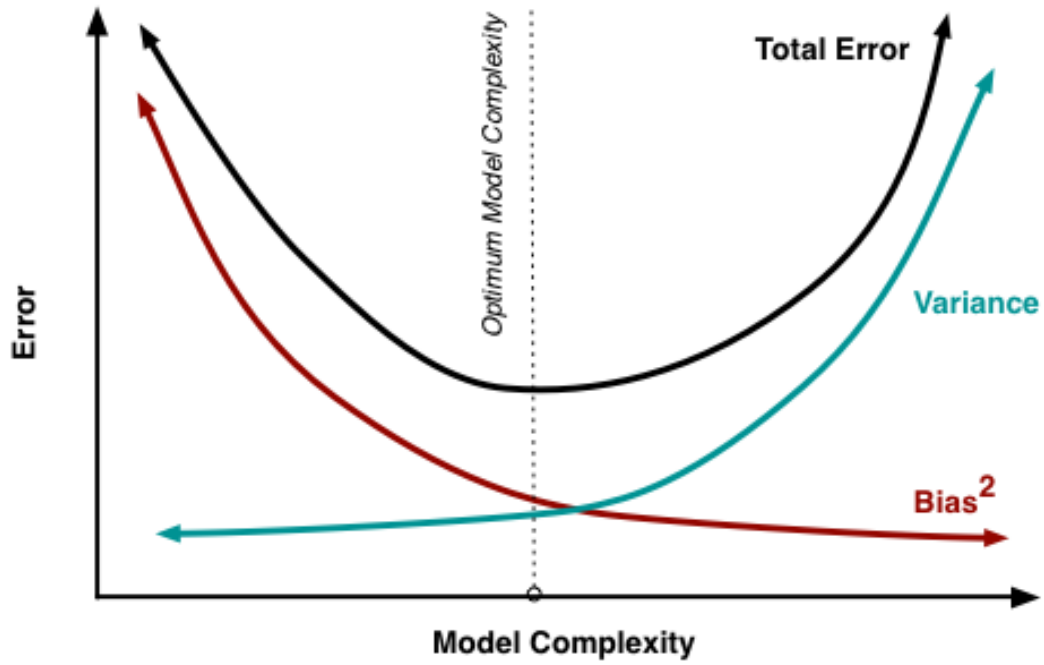
Table 5 ■ Firm value as a function of governance.

Dependent Variable: *Firm q*	(1)	(2)	(3)	(4)	(5)
Property Type q	0.497	0.418	0.403	0.412	0.382
	*(14.33)****	*(7.64)****	*(7.65)****	*(7.16)****	*(10.44)****
EBITDA	0.403	0.456	0.444	0.444	0.163
	*(5.31)****	*(5.17)****	*(5.11)****	*(5.00)****	*(7.21)****
UPREIT	−0.001	−0.006	−0.023	−0.018	
	(0.02)	*(0.09)*	*(0.36)*	*(0.28)*	
Interest Coverage	0.057	0.060	0.043	0.038	−0.004
	(0.74)	*(0.83)*	*(0.63)*	*(0.57)*	*(0.15)*
Mkt Cap	0.127	0.078	0.087	0.096	0.014
	*(2.73)****	*(1.85)**	*(1.96)**	*(2.11)***	*(0.39)*
Excess Comp		−0.002	0.000	−0.002	−0.020
		(0.03)	*(0.01)*	*(0.05)*	*(0.85)*
Instl Ownership		0.053	0.078	0.085	0.101
		(1.00)	*(1.48)*	*(1.50)*	*(2.55)***
Block Ownership			−0.046	−0.041	0.013
			(1.38)	*(1.23)*	*(0.59)*
D&O Ownership			0.106	0.105	0.072
			(1.57)	*(1.55)*	*(2.08)***
Ln(Board Size)				−0.044	−0.097
				(0.77)	*(2.86)****
Outside Board				0.029	0.021
				(0.75)	*(0.93)*
Maryland				−0.026	
				(0.53)	
Fixed Effects?	No	No	No	No	Yes
Observations	882	882	882	882	882
R^2	0.53	0.55	0.56	0.56	0.60
p value from *F* test of null that all governance coefficients are zero		0.61	0.21	0.50	0.00***

Regularization

$$Y = f(X_1, \ldots, X_p) + \varepsilon_i,$$

when predicting Y with a bunch of X's via $f(\cdot)$ we know that...



A fundamental idea in modern statistics (machine leaning) is the use of regularization to explore the bias-variance trade-off

Various flavors: penalized likelihood, priors, smoothing, etc, etc....

Everyone knows...

It is well-known that unmeasured confounders can lead to biased estimates of regression coefficients.

Suppose we're interested in the **treatment effect** of dietary kale intake.

And want to know how effective it is at lowering cholesterol, which is our **outcome variable**.

Unfortunately, we have only observational data (i.e., not a randomized study).

Is it kale or is it gym?

Our bad luck, only gym-rats seem to eat much kale. And exercise is known to lower cholesterol: the "direct" effect is **confounded**.

$$Y_i = \beta_0 + \alpha D_i + \varepsilon_i,$$

Because $\text{cov}(D_i, \varepsilon_i) \neq 0$, we can write

$$Y_i = \beta_0 + \alpha D_i + \omega D_i + \tilde{\varepsilon}.$$

Since $\text{cov}(D_i, \tilde{\varepsilon}_i) = 0$, we mis-estimate α as $\alpha + \omega$.

Easy fix...

The good news is, we can **control** for weekly exercise, X_i, by including it in the regression:

$$Y_i = \beta_0 + \alpha D_i + \beta X_i + \varepsilon_i.$$

This "clears out" the confounding: conditional on X_i, $\text{cov}(D_i, \varepsilon_i) = 0$ and we're good to go.

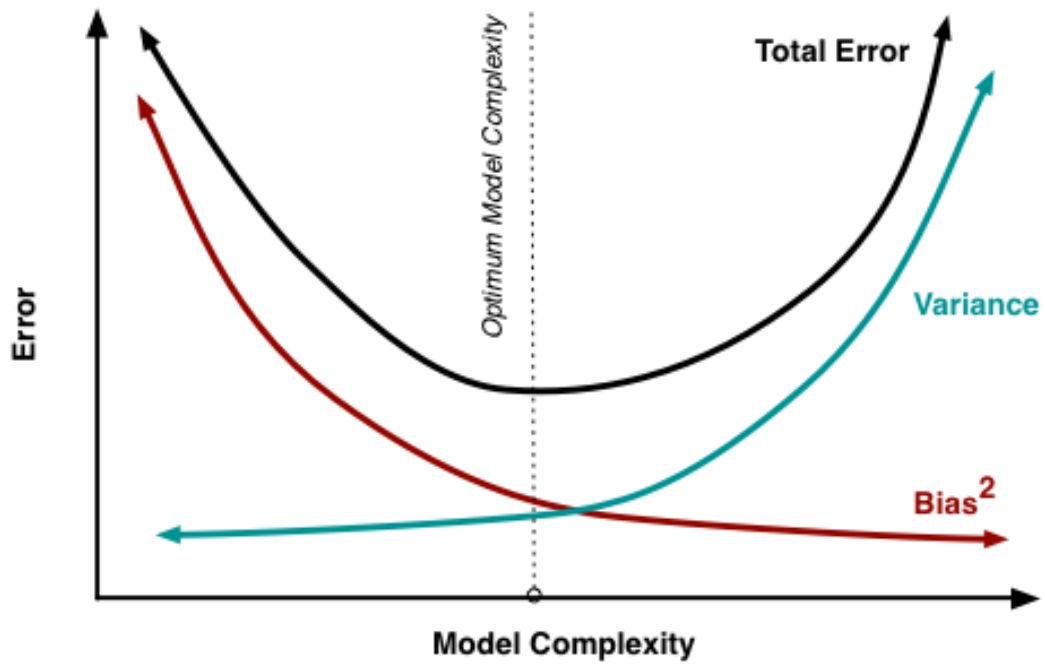
But what if we don't know what we need to control for?

Note: I will assume from this point forward that a subset inside of a large set of variables is enough to identify the treatment effect... i.e. don't ask me about instruments today!!!

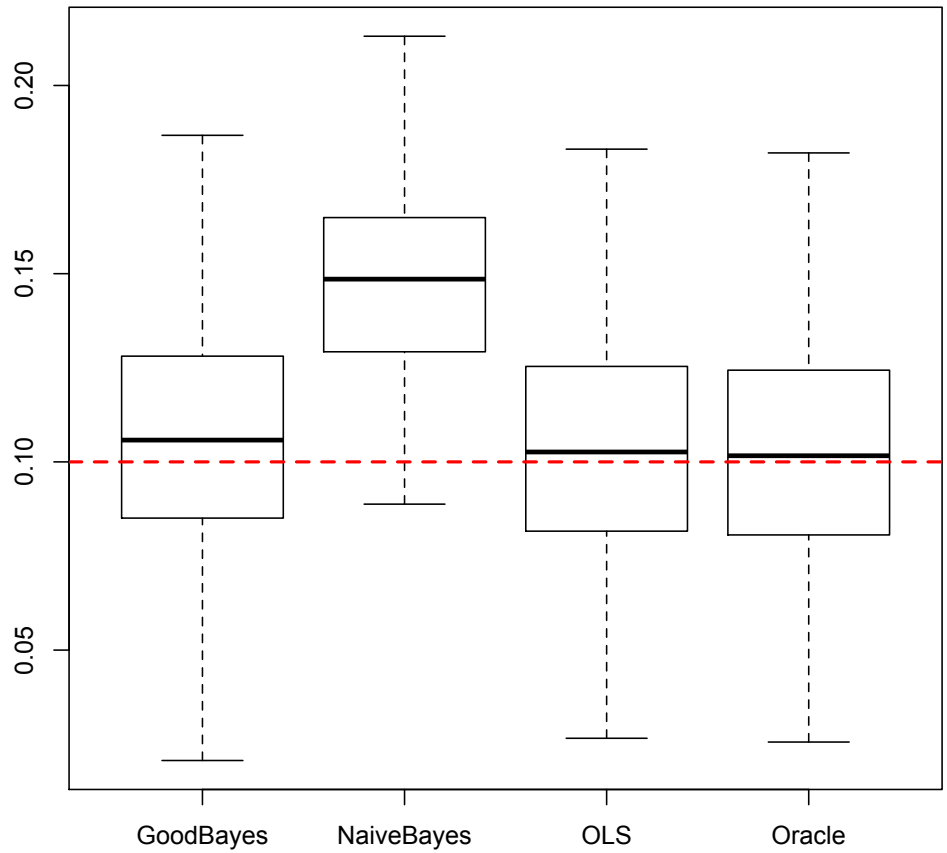
Regularize?

$$Y_i = \beta_0 + \alpha D_i + \mathbf{X}_i'\beta + \varepsilon_i.$$

- Let's regularize using priors... shrinkage prior on β should do it...



Something goes wrong



It turns out that this "obvious" approach is really bad at getting reasonable estimates of the treatment effect α.

Bad bias versus good bias

Assume that:

$$D_i = \mathbf{X}_i^t \gamma + \epsilon_i.$$

Now substitute a shrunk estimate, $\beta - \Delta$, in place of the true (unknown) β vector:

$$Y_i = \alpha D_i + \mathbf{X}_i^t(\beta - \Delta) + [\nu_i + \mathbf{X}_i^t \Delta].$$

This implies that ν_i is taken to be $\nu_i + \mathbf{X}_i \Delta$, which gives

$$\text{cov}(\nu_i + \mathbf{X}_i^t \Delta, \mathbf{X}^t \gamma + \epsilon) \neq 0.$$

Biasing β towards zero biases $\text{cov}(D, \varepsilon)$ *away from* zero!

OLS forces deconfoundedness

It is well-known that in the presence of the "correct" controls,

$$\hat{\varepsilon}^{\text{ols}} \perp D.$$

However, this is not true for regularized estimates

$$\hat{\varepsilon}^{\text{reg}} \not\perp D.$$

We refer to this as **Regularization-induced Confounding**.

The typical parametrization

$$\begin{aligned} \text{Selection Eq.:} \quad & D = \mathbf{X}^t\gamma + \epsilon, && \epsilon \sim N(0, \sigma_\epsilon^2), \\ \text{Response Eq.:} \quad & Y = \alpha D + \mathbf{X}^t\beta + \nu, && \nu \sim N(0, \sigma_\nu^2). \end{aligned}$$

These equations correspond to the factorization of the joint distribution

$$f(Y, D \mid \gamma, \beta, \sigma_\epsilon, \sigma_\nu) = f(Y \mid D, \beta, \sigma_\epsilon) f(D \mid \gamma, \sigma_\nu).$$

This factorization implies a complete separation of the parameter sets: independent priors on the regression parameters

$$\pi(\beta, \gamma, \alpha) = \pi(\beta)\pi(\gamma)\pi(\alpha)$$

imply that only the response equation is used in estimating β and α.

Our reparametrization: a latent error approach

We reparametrize as

$$\begin{pmatrix} \alpha \\ \beta + \alpha\gamma \\ \gamma \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ \beta_d \\ \beta_c \end{pmatrix}.$$

which gives the new equations

$$\begin{aligned} \text{Selection Eq.:} \quad & D = \mathbf{X}^t\beta_c + \epsilon, && \epsilon \sim N(0, \sigma_\epsilon^2), \\ \text{Response Eq.:} \quad & Y = \alpha(D - \mathbf{X}^t\beta_c) + \mathbf{X}^t\beta_d + \nu, && \nu \sim N(0, \sigma_\nu^2). \end{aligned}$$

We can now shrink β_d and β_c with impunity!

Simulation study

$$\begin{aligned}
&\text{Selection Eq.:} && D = \mathbf{X}^t\beta_c + \epsilon, && \epsilon \sim N(0, \sigma_\epsilon^2),\\
&\text{Response Eq.:} && Y = \alpha(D - \mathbf{X}^t\beta_c) + \mathbf{X}^t\beta_d + \nu, && \nu \sim N(0, \sigma_\nu^2).
\end{aligned}$$

Set $var(D) = var(Y) = 1$ and center and scale the columns of $\mathbf{X}$.

Define the ℓ_2 norms of the confounding and direct effects as $\rho^2 = \|\beta_c\|_2^2$ and $\phi^2 = \|\beta_d\|_2^2$ so that

$$\begin{aligned}
var(D) &= \rho^2 + \sigma_\epsilon^2\\
var(Y) &= \kappa^2 + \phi^2 + \sigma_\nu^2,
\end{aligned}$$

with $\sigma_\epsilon^2 = 1 - \rho^2$ and $\sigma_\nu^2 = 1 - \alpha^2(1 - \rho^2) - \phi^2$ and $\kappa^2 = \alpha^2(1 - \rho^2)$.

ρ^2		Bias	Coverage	I.L.	MSE
0.1	New Approach	-0.0032	0.943	0.2357	0.0037
	OLS	-0.0016	0.951	0.2477	0.004
	Naive Regularization	-0.0112	0.895	0.2089	0.0037
	Oracle OLS	0.0023	0.946	0.2173	0.0031
0.3	New Approach	-0.0047	0.95	0.2751	0.0047
	OLS	-0.0018	0.951	0.2808	0.0052
	Naive Regularization	-0.0355	0.848	0.2293	0.0057
	Oracle OLS	0.0026	0.946	0.2464	0.004
0.5	New Approach	-3e-04	0.963	0.3345	0.0066
	OLS	-0.0022	0.951	0.3323	0.0072
	Naive Regularization	-0.0768	0.746	0.2631	0.012
	Oracle OLS	0.0031	0.946	0.2915	0.0056
0.7	New Approach	0.0084	0.964	0.4374	0.0113
	OLS	0.0024	0.944	0.4303	0.0123
	Naive Regularization	-0.1559	0.543	0.3292	0.0346
	Oracle OLS	0.004	0.946	0.3764	0.0093
0.9	New Approach	-0.004	0.972	0.7403	0.0292
	OLS	0.0045	0.954	0.7469	0.0351
	Naive Regularization	-0.4482	0.231	0.4779	0.2391
	Oracle OLS	0.0069	0.946	0.6519	0.0278

Table: $\mathbf{n = 100, p = 30, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

ρ^2		Bias	Coverage	I.L.	MSE
0.1	New Approach	0.0082	0.918	0.3632	0.0105
	OLS	-0.0017	0.944	0.4785	0.0144
	Naive Regularization	-0.0068	0.835	0.2957	0.0097
	Oracle OLS	-0.001	0.952	0.3235	0.0065
0.3	New Approach	-1e-04	0.94	0.4203	0.0128
	OLS	-0.002	0.944	0.5425	0.0186
	Naive Regularization	-0.035	0.837	0.3191	0.0126
	Oracle OLS	-0.0011	0.952	0.3668	0.0084
0.5	New Approach	-0.0047	0.93	0.5183	0.0196
	OLS	-0.0023	0.944	0.6419	0.026
	Naive Regularization	-0.0869	0.738	0.3555	0.0222
	Oracle OLS	-0.0014	0.952	0.434	0.0117
0.7	New Approach	0.0056	0.937	0.6926	0.0341
	OLS	0.0046	0.934	0.8204	0.0478
	Naive Regularization	-0.189	0.539	0.4033	0.0565
	Oracle OLS	-0.0018	0.952	0.5604	0.0195
0.9	New Approach	-0.0772	0.959	1.1572	0.0804
	OLS	-0.0156	0.931	1.4347	0.1402
	Naive Regularization	-0.5419	0.102	0.4868	0.3297
	Oracle OLS	-0.003	0.952	0.9706	0.0585

Table: $\mathbf{n = 50, p = 30, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

Empirical example: Levitt abortion reanalysis

According to "Freakonomics":

- unwanted children are more likely to grow up to be criminals,
- therefore legalized abortion, which leads to fewer unwanted children, leads to lower levels of crime in society.

To investigate, they conduct three analyses, one each for three different types of crime: violent crime, property crime, and murders.

Donohue III and Levitt data

Y is per capita crime rates (violent crime, property crime, and murders) by state, from 1985–1997, and D, is the "effective" abortion rate.

The control variables, $\mathbf{X}$, are:

- prisoners per capita (log),
- police per capita (log),
- state unemployment rate,
- state income per capita (log),
- percent of population below the poverty line,
- generosity of AFDC (lagged by fifteen years),
- concealed weapons law,
- beer consumption per capita.

Including state and year dummy variables brings the total number of control variables to $p = 66$ (with $n = 624$).

Replication

	Property Crime		Violent Crime		Murder	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
OLS	-0.110	-0.072	-0.171	-0.090	-0.221	-0.040
Our way	-0.113	-0.073	-0.182	-0.098	-0.222	-0.039
naive	-0.075	-0.010	0.079	0.301	-0.186	0.085

An augmented control set

Our expanded model includes the following additional control variables:

- interactions between the original eight controls and year,
- interactions between the original eight controls and year squared,
- interactions between state effects and year,
- interactions between state effects and year squared.

When allowing for this degree of flexibility, estimation becomes quite challenging, with just $n = 624$ observations and $p = 176$ control variables.

Augmented analysis results

	Property Crime		Violent Crime		Murder	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
OLS	-0.226	0.019	-0.374	0.336	-0.125	1.763
Our way	-0.038	0.014	-0.114	0.053	-0.081	0.279
naive	0.007	0.129	0.011	0.412	-0.227	0.116

Nonlinear/heterogeneous regressions for deconfounding

Consider the nonlinear model

$$Y_i = f(X_i, D_i) + \varepsilon_i.$$

Now our deconfoundedness condition is stronger

$$D_i \perp \varepsilon_i \mid X_i.$$

And our causal estimate is more general

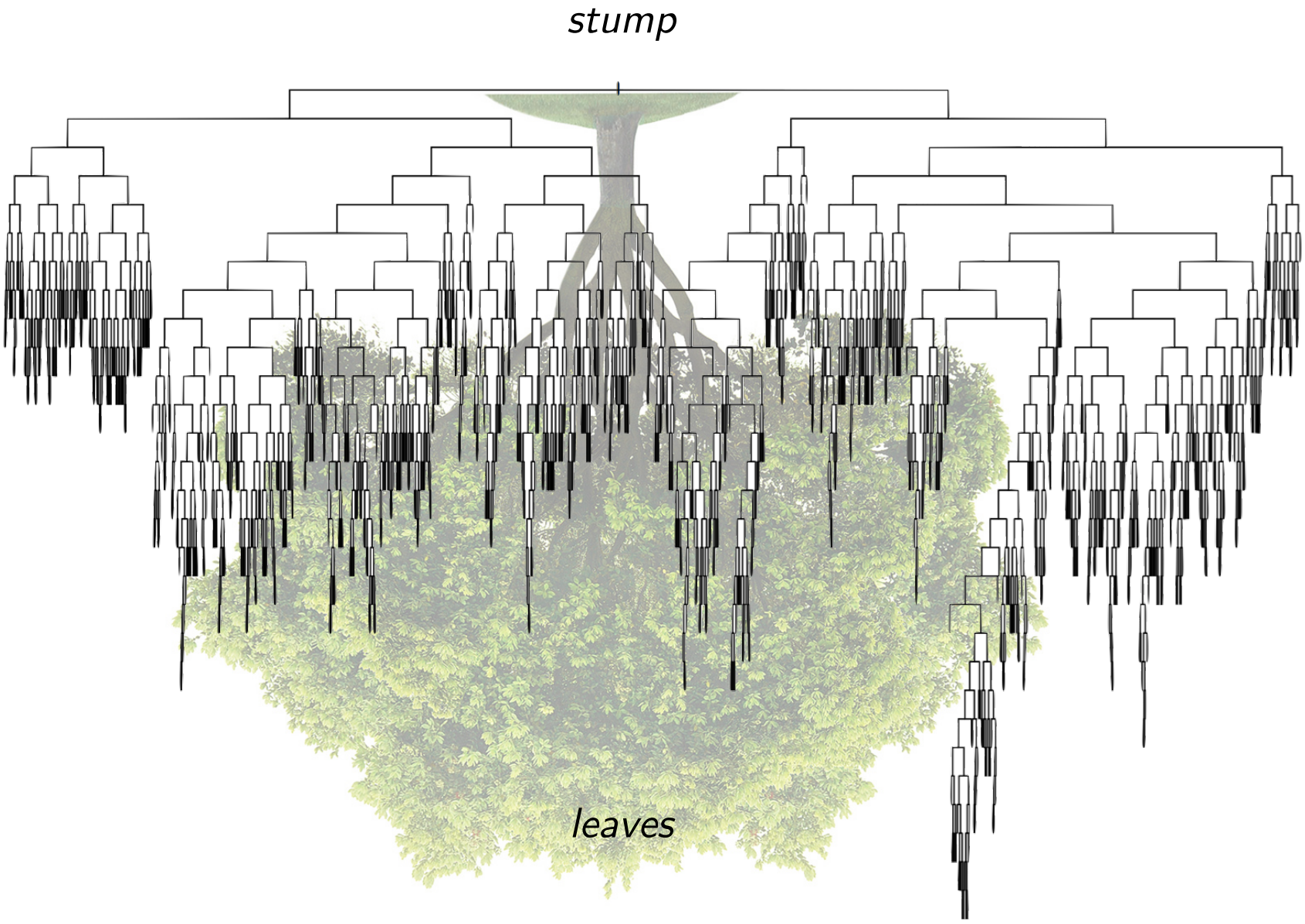
$$\alpha(X_i, D_i) = \frac{\partial f(X_i, D_i)}{\partial D_i}.$$

Here we have no option but to regularize!

How about using our favorite tree model

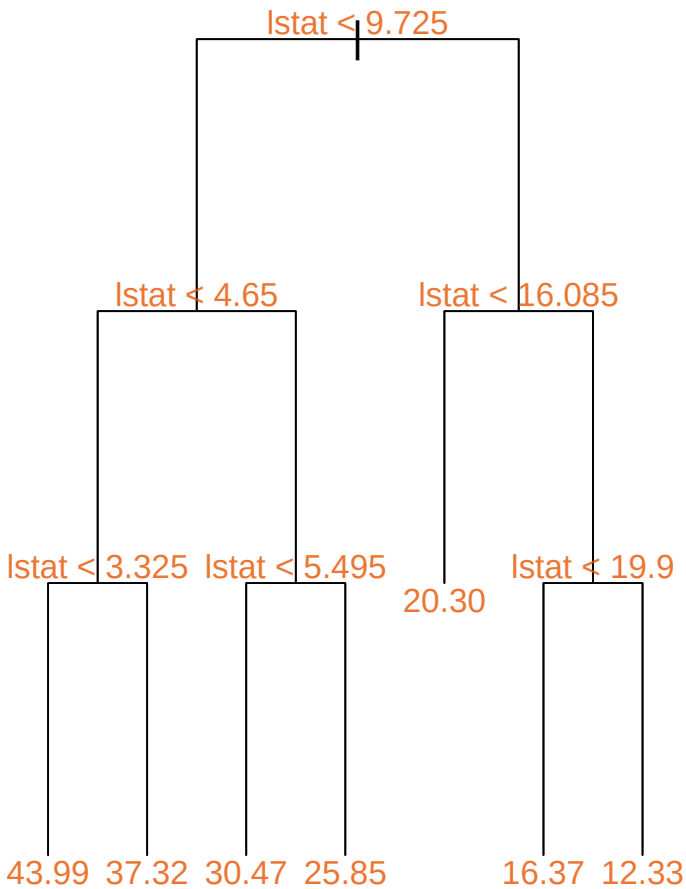
Trees

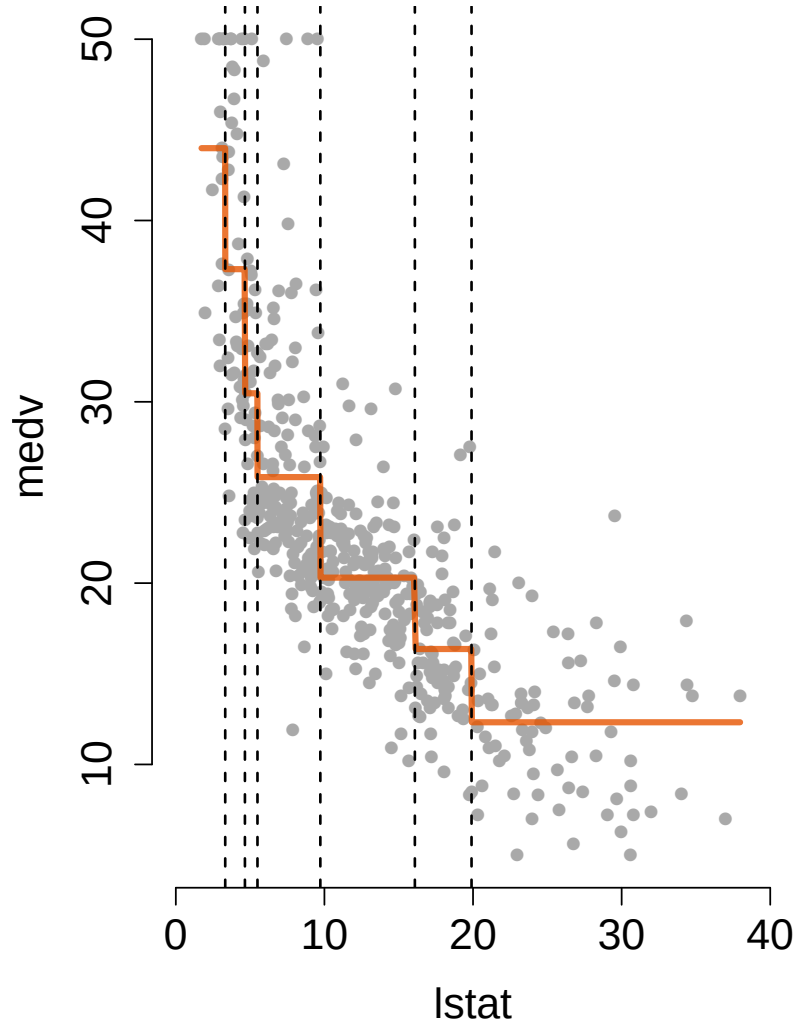
Trees



Basic structure - regression trees

Follow rules down tree to come up with prediction. The resulting $f(\cdot)$ is a *step function*! Each region is *average* of training data.





Regularization-induced confounding (again)

Nothing in the standard "tree" likelihood that encourages

$$D_i \perp \varepsilon_i \mid X_i.$$

in that the likelihood takes no heed of the relationship between $Y_i - \mu_{\tau(X_i)}$ and D_i.

In the linear setting a **joint propensity-response model** solves this problem:

$$\begin{aligned} &\text{Selection Eq.:} && D = \mathbf{X}^t\beta_c + \epsilon, && \epsilon \sim N(0, \sigma_\epsilon^2), \\ &\text{Response Eq.:} && Y = \alpha(D - \mathbf{X}^t\beta_c) + \mathbf{X}^t\beta_d + \nu, && \nu \sim N(0, \sigma_\nu^2). \end{aligned}$$

We will apply the same strategy to the regression tree setting.

A bivariate treed linear model

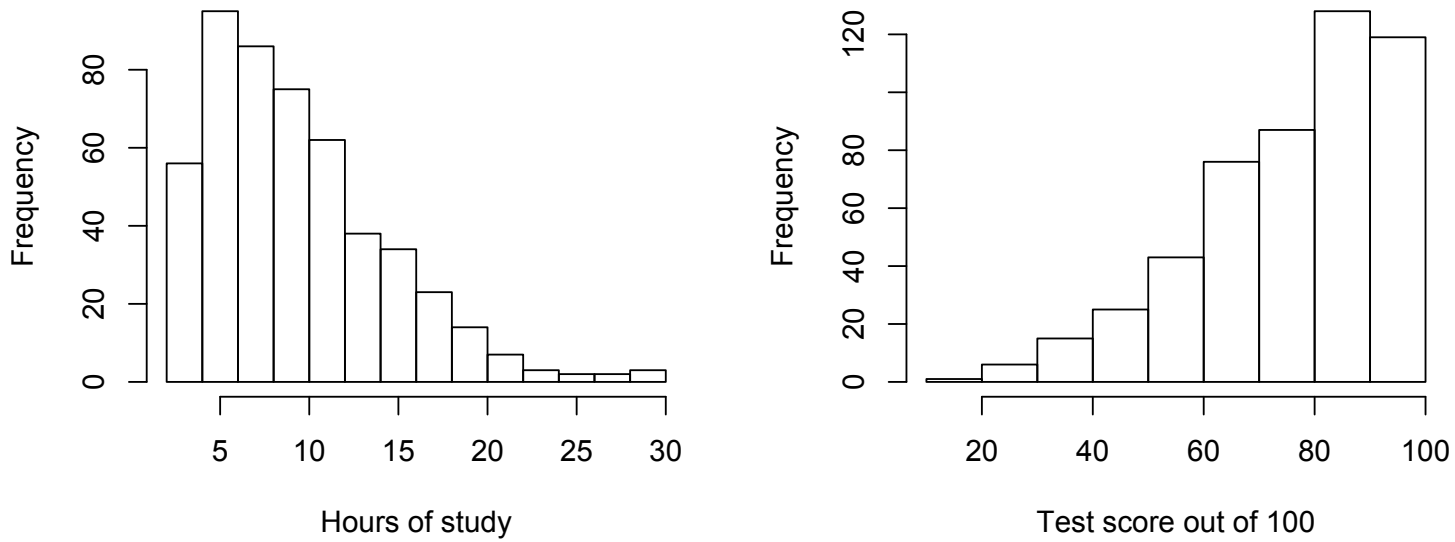
We again have a regression tree $\tau(X_i)$, but at each leaf we have the joint likelihood model $f(Y \mid D)f(D)$

$$Y_i = \alpha^0_{\tau(X_i)} + \alpha_{\tau(X_i)} D_i + \varepsilon_i$$
$$D_i = \mu_{\tau(X_i)} + \epsilon_i.$$

- and the likelihood asserts that $\varepsilon_i \perp \epsilon_i$.
- In each leaf, the evaluated counterfactual is based on independent variation in dosage D_i.

Simulated example

Consider estimating the treatment effect of hours of tutoring D_i on a certain test score Y_i.



Suppose that we control for IQ (X_1) and family income (X_2), as well as fifteen additional attributes of the individual test taker.

Data generating process

The selection process is

$$
\begin{aligned}
D_i &= \exp\left\{\arctan(-X_1/3) + \arctan(X_2) + \log(8) + 0.3\epsilon_i\right\} \\
\epsilon_i &\sim \mathrm{N}(0,1).
\end{aligned}
$$

The response process is

$$
\begin{aligned}
U_i &= (5\arctan(X_2 - 3) + \pi/2) + 1)\arctan(Z), \\
S_i &= (\arctan(U_i) + \pi/2)/\pi, \\
Y_i &= 100 \times \Phi(\Phi^{-1}(S_i) + 0.5\varepsilon_i), \\
\varepsilon_i &\sim \mathrm{N}(0,1).
\end{aligned}
$$

These aren't as crazy as they look.

A horse race

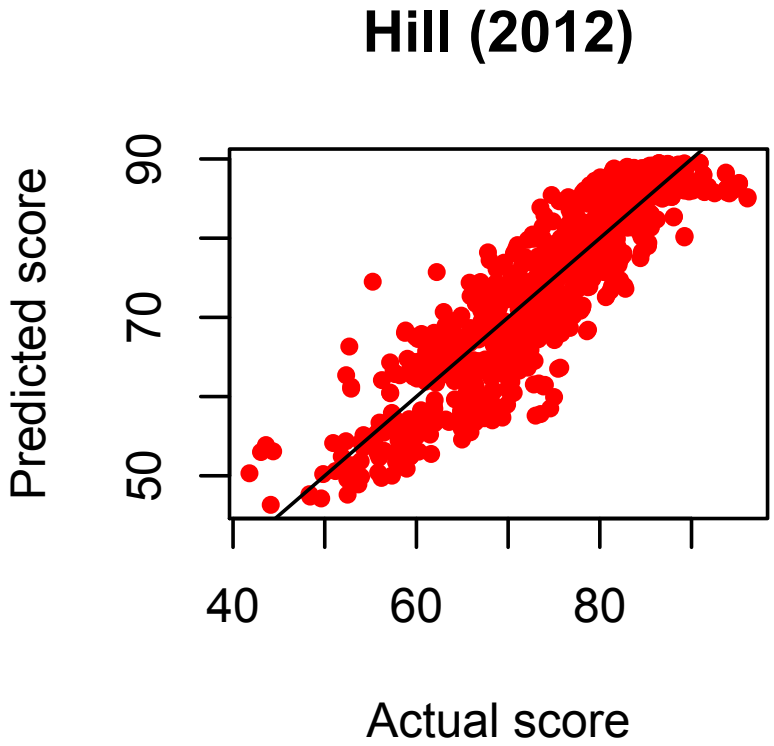
We compare four models ($n = 500$, $p = 17$)

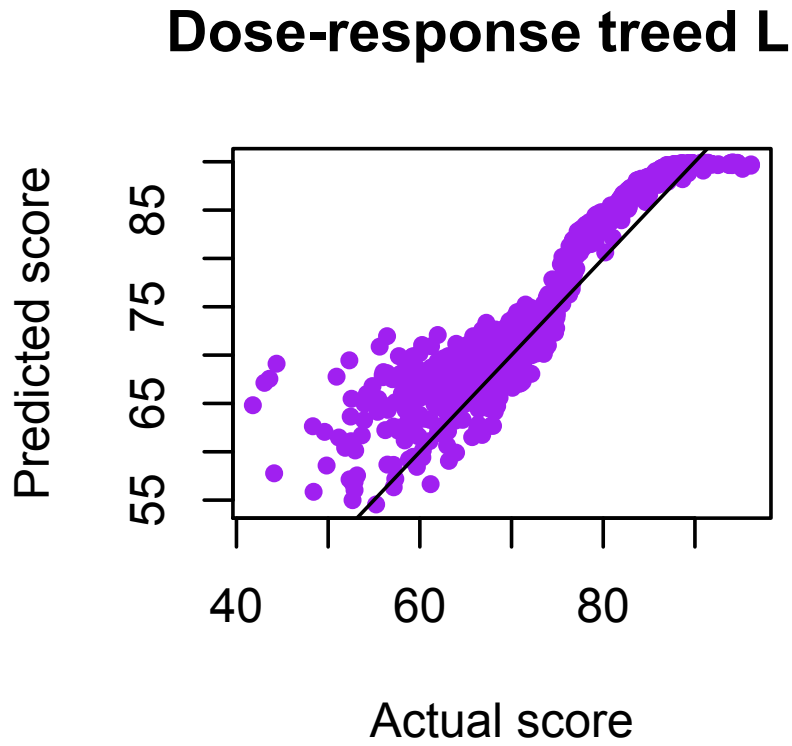
1. BART *a la* Hill (2012), (naive approach)

2. Bivariate dose-response treed linear model.

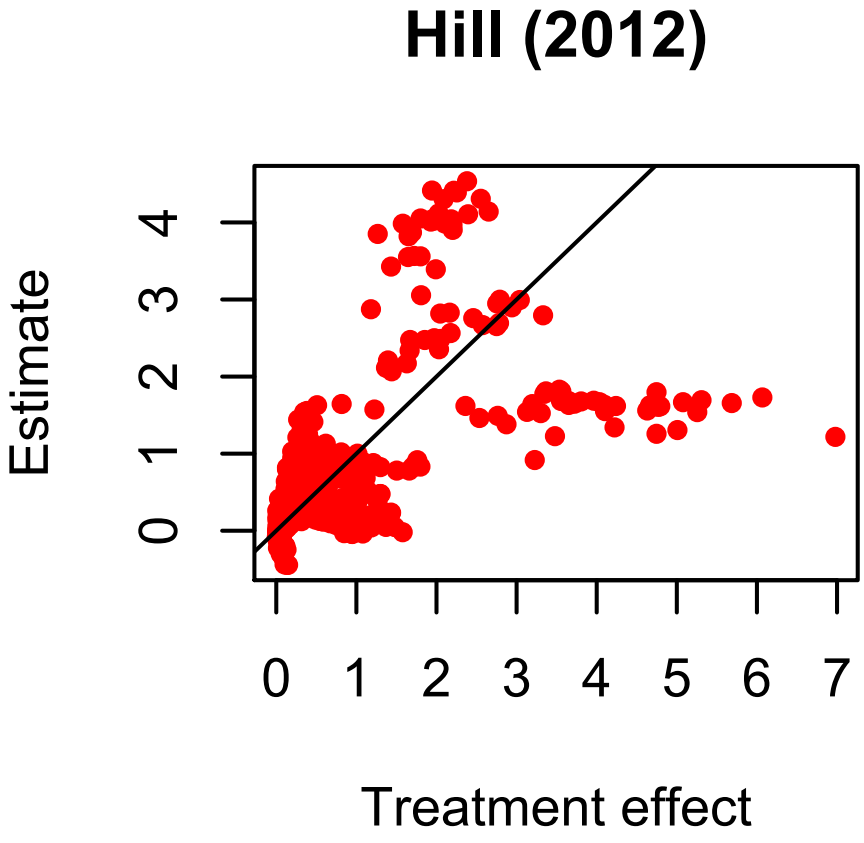
Our estimand will be $\mathrm{E}(Y_i \mid X_i, D_i = d_i + 1) - \mathrm{E}(Y_i \mid X_i, D_i = d_i)$: the individual causal effect of one additional hour of test prep.

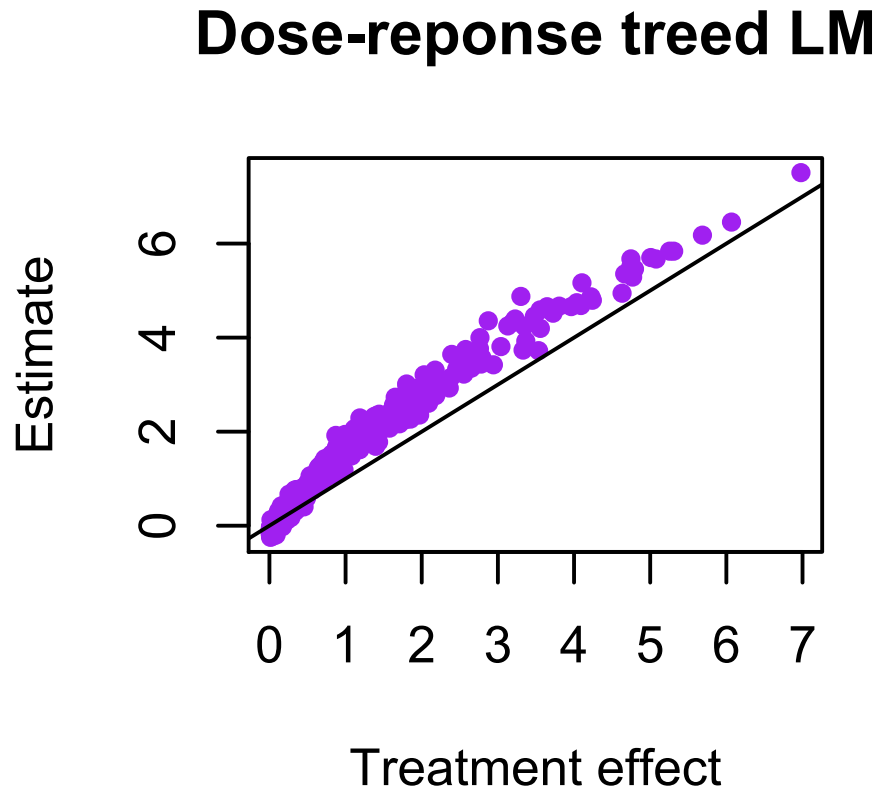
Prediction results





Treatment effect estimation results





Numbers

Model	Prediction RMSE	Estimation RMSE
Hill (2012)	4.62	0.98
DR treed LM	4.31	0.46

Summary

- Regularization-induced confounding is a thing that happens.
- Explicitly modeling the treatment allows regularization to be imposed robustly.
- This opens the door to the effective use of lots of powerful tools for the estimation of causal effects!
- Lot's to do before a stata function is available... :)