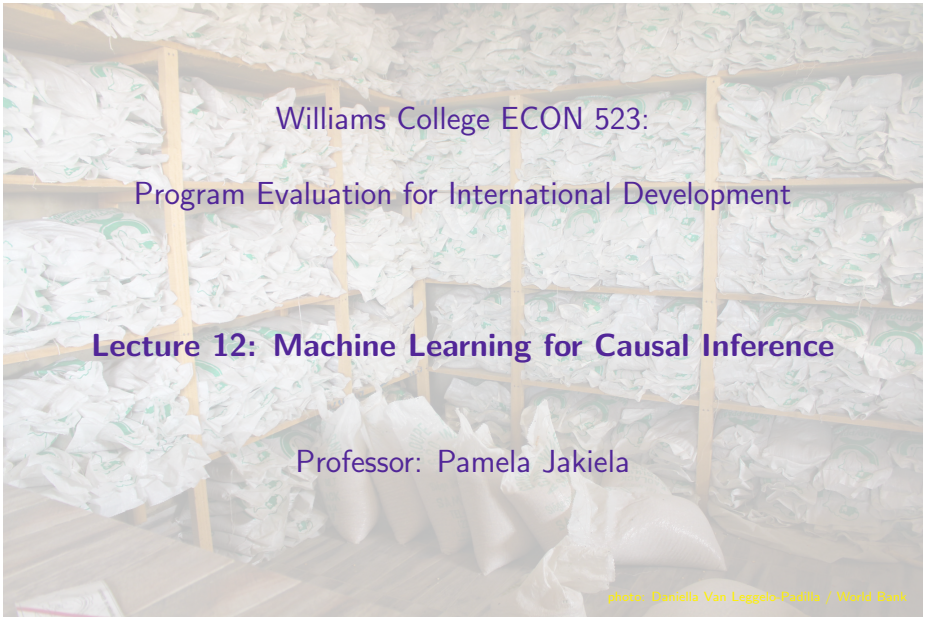


Williams College ECON 523:

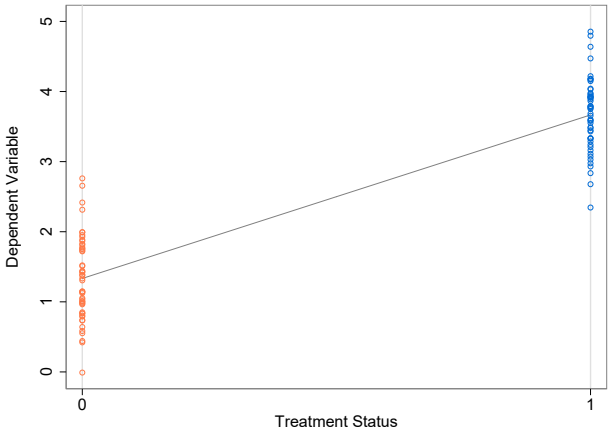
Program Evaluation for International Development

Lecture 12: Machine Learning for Causal Inference

Professor: Pamela Jakiela

photo: Daniella Van Leggelo-Padilla / World Bank

Basic Regression Equation: $Y_i = \alpha + \beta D_i + \varepsilon_i$



RCT Regression Specification with Controls

More typical regression specification:

$$Y_{1,i} = \alpha + \beta D_i + \delta X_{0,i} + \gamma Y_{0,i} + \kappa_{strata} + \varepsilon_i$$

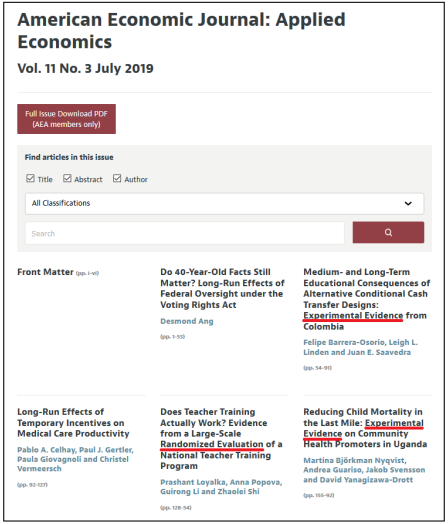
We typically include these controls:

- Dummies for randomization strata (κ_{strata})
- Selected baseline covariates that are not balanced across treatments*
- Baseline covariates that predict the outcome
  - Baseline values of outcome variables are (sometimes) most important (ANCOVA)

We do not want to include:

- Controls that could be impacted by treatment ("bad controls" problem)

"You Don't Have to Take My Word For It"



American Economic Journal: Applied Economics

Vol. 11 No. 3 July 2019

Full Issue Download PDF (AEA members only)

Find articles in this issue

Title Abstract Author

All Classifications

Search

Front Matter (pp. i-vi)

Do 40-Year-Old Facts Still Matter? Long-Run Effects of Federal Oversight under the Voting Rights Act
Desmond Ang
(pp. 1-33)

Medium- and Long-Term Educational Consequences of Alternative Conditional Cash Transfer Designs: Experimental Evidence from Colombia
Felipe Barrera-Osorio, Leigh L. Linden and Juan E. Saavedra
(pp. 54-91)

Long-Run Effects of Temporary Incentives on Medical Care Productivity
Pablo A. Celhay, Paul J. Gertler, Paula Giovagnoli and Christel Vermeersch
(pp. 92-127)

Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program
Prashant Loyalka, Anna Popova, Guirong Li and Zhaolei Shi
(pp. 128-54)

Reducing Child Mortality in the Last Mile: Experimental Evidence on Community Health Promoters in Uganda
Martina Björkman Nyqvist, Andrea Guariso, Jakob Svensson and David Yanagizawa-Drott
(pp. 155-92)

"You Don't Have to Take My Word For It"

American Economic Journal: Applied Economics 2019, 11(3): 128–154
https://doi.org/10.1257/app.20170226

Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program[†]

By Prashant Loyalka, Anna Popova, Guirong Li, and Zhaolei Shi[*]

Despite massive investments in teacher professional development (PD) programs in developing countries, there is little evidence on their effectiveness. We present results of a large-scale, randomized evaluation of a national PD program in China in which teachers were randomized to receive PD; PD plus follow-up; PD plus evaluation of the command of PD content; or no PD. Precise estimates indicate PD and associated interventions failed to improve teacher and student outcomes after one year. A detailed analysis of the causal chain shows teachers find PD content to be overly theoretical, and PD delivery too rote and passive, to be useful. (*JEL* I21, I28, J24, J45, O15, P36)

"You Don't Have to Take My Word For It"

We estimate the ATEs using the following ordinary least squares regression model:[22]

$$Y_{ij} = \alpha_0 + \alpha_1 D_j + X_{ij}\alpha + \tau_k + \varepsilon_{ij}, \quad (1)$$

where Y_{ij} is the outcome of interest measured at endline for student i in school j; D_j is one or more dummies indicating the treatment assignment of school j; X_{ij} is a vector of baseline control variables; and τ_k is a set of block fixed effects. In all specifications, X_{ij} includes the baseline value of the dependent variable whenever this is available. We also estimate treatment effects with an expanded set of baseline controls (we call these our "covariate-adjusted" regressions). For student-level outcomes, this expanded set of controls includes student age, student gender, parent educational attainment, a household asset index, class size, teacher gender, teacher age, teacher experience, teacher education level, a teacher certification dummy, a teacher major in math dummy, and teacher rank. For outcomes measured at the teacher level, student controls are omitted.

We Want to Include Controls that Predict T_i or Y_i

Controls should be orthogonal to treatment status (because we randomized):

- In practice, a control might be correlated with treatment
  - In small(ish) samples, we may see some differences in observables/covariates between treatment and comparison groups (especially as the number of covariates increases)
- Can be important to show estimated impacts are "robust" to inclusion of covariates
- Adding in one imbalanced covariate can undo randomization (remember, we are now regressing residual of Y on residual of D)
  - This is (one of the reasons) why **stratification** is desirable

Baseline covariates are orthogonal to treatment, so they should not impact coefficient

- Controls help if they explain residual variation in Y

We Want to Include Controls that Predict T_i or Y_i

$$MDE = \left(t_{1-\kappa} + t_{\alpha/2}\right) \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\tilde{\sigma}^2}{N}} \sqrt{1 + (n_g - 1)\,\tilde{\rho}}$$

Best Subset Selection

A **best subset selection** algorithm:

- For each $k = 1, 2, \ldots, p$
  - Fit all models containing exactly k covariates
  - Identify the "best" in terms of R^2
- Given a frontier of best subsets (conditional on k), need to choose optimal k
  - Need to address the fact that R^2 always increases with k
  - Multiple approaches: adjusted R^2, cross-validation, etc.

When p is large, best subset selection is not feasible (too many combinations of variables)

- Forward and backward subset selection may work poorly when covariates are correlated

Best Subset Selection

In OLS, we seek to minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

Best subset selection can be expressed as: choose β to minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p} I\left(\beta_j \neq 0\right) \leq s$$

where s is the number of regressors/covariates/predictors/features included in the model

$\Rightarrow$ But we solve it algorithmically, not analytically

$\Rightarrow$ When p is large, finding the best subset is hard

What Is Machine Learning?

A set of extensions to the standard econometric/statistical toolkit aimed at improving predictive accuracy, particularly when data sets are sparse (many variables, most are garbage)

- Like subset selection, lasso is an extension of OLS, useful for covariate selection
- Other ML methods are increasingly being used by economists to capture treatment effect heterogeneity, to identify latent groupings in data (e.g. competitors), and to analyze text

Machine learning introduces new tools, relabels existing tools

- Main focus is on predicting Y, not testing hypotheses about causal impact of T

  $\Rightarrow$ ML "results" about β may not be robust or causally identified

- Vocabulary: test vs. training data/sample/examples, features, train a model

Shrinkage Operators: Machine Learning Extensions to OLS

Machine learning **shrinkage operators** (ridge regression, lasso) extend OLS to better predict Y

- Basic idea is to fully "kitchen sink" our regressions while proactively correcting for potential over-fitting, allowing us to leverage information from more covariates effectively

Lasso is attractive because it identifies a subset of Xs that are most effective predictors of Y

Can We Improve on OLS?

A standard linear model may not be the best way to predict Y:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Can we improve on OLS?

- When p is large relative to N, OLS is prone to over-fitting
- OLS explains both structural and spurious relationships in data

Like best subset selection, shrinkage operators minimize RSS subject to an additional constraint

$$\min_\beta \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } f(\beta) \leq s$$

Shrinkage Operators: Ridge Regression

Ridge regression solves a closely related minimization problem:

$$\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

or, equivalently,

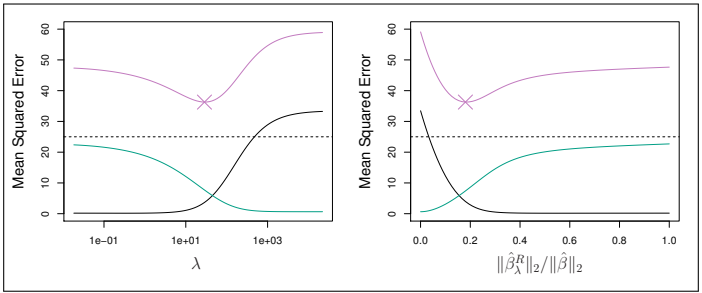
$$\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

for some **tuning parameter** $\lambda \geq 0$

Ridge regression shrinks OLS coefficients toward zero

- Shrinkage is more or less proportional, so ridge regression does not identify a subset of regressors to include in the regression model (it just down-weights some relative to others)

OLS is BLUE, But Ridge Regression (Sometimes) Has Lower MSE



Source: James et al. (2021)

Gauss-Markov Theorem: OLS is the **best linear unbiased estimator** (BLUE) of Y

- Ridge regression is biased (black line), but has lower variance relative to the true underlying β (green line) and can therefore achieve lower MSE (pink line) for some λs

Shrinkage Operators: Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) seeks to minimize:

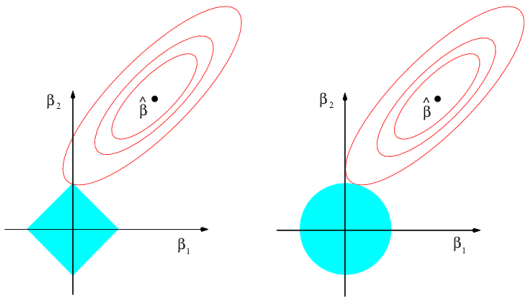
$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

for some **tuning parameter** $\lambda \geq 0$

Lasso combines benefits of subset selection, ridge regression

- Less computationally intensive than subset selection
- Sets some coefficients to 0 $\rightarrow$ identifies parsimonious model

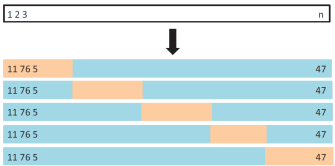
Lasso Sets Some Coefficients to Zero



Source: James et al. (2021)

The lasso constraint region has sharp corners $\Rightarrow$ some coefficients set to 0

Three Approaches to Choosing λ (1/3)



Source: Witten et al. (2023)

k-fold cross-validation (default with Stata's `lasso` command):

- Randomly sort observations in k groups
- For each group k, estimate lasso on on rest of sample and predict MSE using observations in k (the hold-out sample); average to get MSE(λ)
- Iterate over λ values to choose λ that minimizes MSE

Three Approaches to Choosing λ (2/3)

Bayesian Information Criterion (BIC): function of n, RSS, number of parameters in the model

$$BIC = \frac{1}{n}\left(RSS + \log(n) d \hat{\sigma}^2\right)$$

where d is the number of covariates and $\hat{\sigma}^2$ is an unbiased estimate of the variance

- One of several examples of approaches based on (essentially) adjusted R^2
- `bic` is the only one of these options available from Stata's `lasso` command

Three Approaches to Choosing λ (3/3)

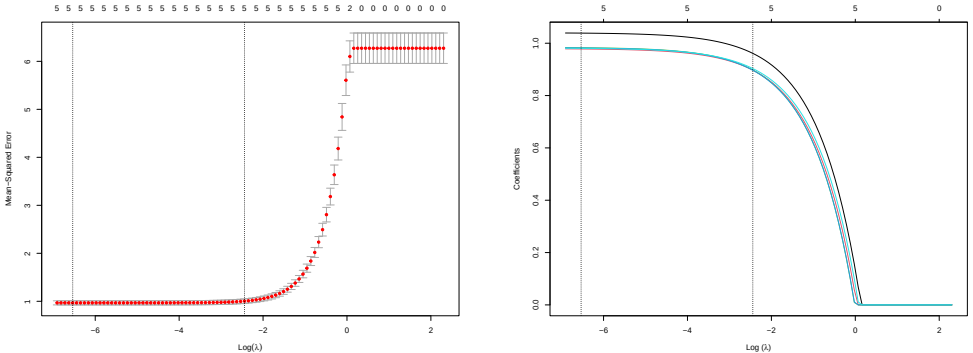
Belloni et al. (2012, 2013, 2014), etc.: alternative approach to choosing λ

- Chooses λ iteratively based on data, penalty can vary across covariates
- Errs on the side of choosing fewer controls to avoid over-fitting
- Allows for heteroskedasticity

Three approaches may generate very different sets of controls

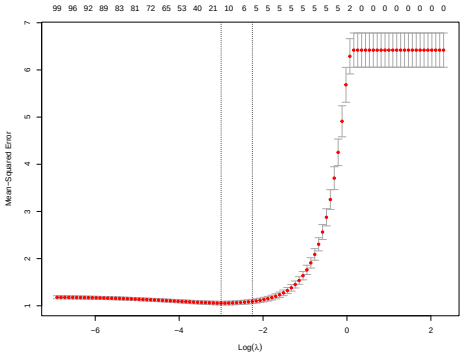
- Costs of too many/too few may vary across empirical contexts

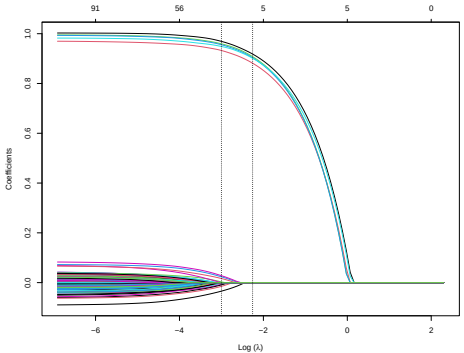
Lasso in Simulated Data: $N = 1000$, $K = 5$



data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 5$, $\varepsilon \sim N(0,1)$, $N = 1000$, $K = 5$

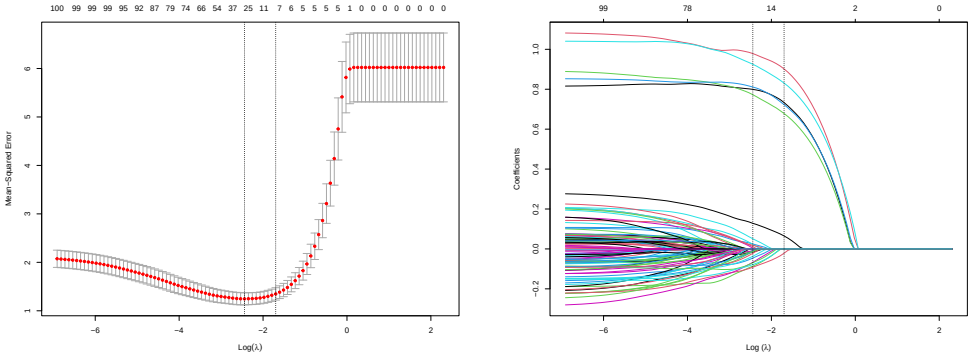
Lasso in Simulated Data: $N = 1000$, $K = 100$





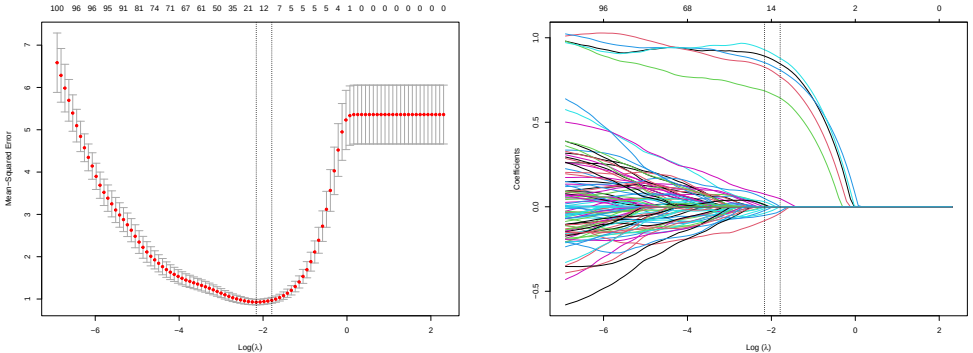
data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 1000$, $K = 100$

Lasso in Simulated Data: $N = 200$, $K = 100$



data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 200$, $K = 100$

Lasso in Simulated Data: $N = 120$, $K = 100$



data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 120$, $K = 100$

Using Lasso to Choose Covariates in an RCT

Conceptually, we want to control for covariates that:

- Predict the outcome conditional on treatment (to increase power/precision)
- Predict treatment (i.e. are slightly imbalanced by chance, which only matters ex post)

Over-fitting, noisy controls can make our treatment effect estimates **less** precise

- Collecting a larger number of covariates also costs more (implying a lower sample size?)

Post-Double-Selection (PDS) Lasso

Norm is to use **post-double-selection** (PDS) lasso to select covariates expost:

- Use lasso to identify baseline covariates that predict treatment
- Use lasso to identify baseline covariates that predict Y (**not** controlling for T)
- Include all of the above in regression estimates of treatment effects
  - Also include **amelioration set** including strata fixed effects, baseline Y, etc.

Including predictors of T matters a lot when treatment is not randomly assigned, less in RCTs

- Addresses any concerns about imbalance across treatment, control groups

PDS lasso allows for standard inference procedures

PDS Lasso: Malaria Data Example

Variable	Description
act_any	RCT arm: assigned ACT any subsidized price (less than 500 KSh)
c_act	Episode care: Treated with ACT
b_h_edu	Baseline: years of education of household head
b_knowledge	Baseline: Knows only mosquitoes transmit malaria
b_hh_size	Baseline: Number of household members
b_acres	Baseline: Acres of land owned by household head
b_dist_km	Baseline: Distance (km) from household to study chemist
b_h_age_imputed	Baseline: Age of head (missing replaced by sample mean)
b_h_age_missing	Baseline: Age of head missing

Preparing the Data

Variable	Obs.	Mean	SD	Min.	Max.
act_any	575	0.7391304	0.4394912	0	1
c_act	575	0.3408696	0.4744143	0	1
b_h_edu	572	5.300699	3.940068	0	16
b_knowledge	575	0.5391304	0.4989005	0	1
b_hh_size	575	5.466087	2.48658	1	14
b_acres	462	2.270779	2.697221	0	34
b_dist_km	574	1.667154	.9307161	.0332327	3.982576
b_h_age_imputed	575	39.2537	15.3422	17	88
b_h_age_missing	575	0.0504348	0.2190309	0	1

Using Stata's lasso Command: Data-Driven Penalty

```
lasso linear act_any b_* miss_*, selection(plugin)
```

...

```
display "‘e(othervars_sel)’"
```

```
b_h_edu b_h_age_imputed b_h_age_missing
```

```
lasso linear act_any b_*, selection(plugin)
display "‘e(othervars_sel)’"
```

```
miss_b_dist_km
```

Using Stata's `lasso` Command: Data-Driven Penalty

```
lasso linear act_any b_* miss_*, selection(plugin)
```

...

```
display "`e(othervars_sel)'"
```

b_h_edu b_h_age_imputed b_h_age_missing

```
lasso linear act_any b_*, selection(plugin)
display "`e(othervars_sel)'"
```

miss_b_dist_km *(will often be "no covariates selected")*

Using Stata's lasso Command: CV-Selected Penalty

```
lasso linear act_any b_* miss_*, selection(cv)
```

...

```
display "‘e(othervars_sel)’"
```

```
b_h_edu b_knowledge b_hh_size b_acres b_dist_km b_h_age_imputed ...
...b_h_age_missing miss_b_h_edu miss_b_acres
```

```
lasso linear act_any b_*, selection(cv)
display "‘e(othervars_sel)’"
b_acres miss_b_acres
```

Using Stata's `lasso` Command: Results

	OLS	**PDS:DD**	**PDS:CV**
	(1)	(2)	(3)
Treatment	0.190	0.186	0.190
	(0.040)	(0.039)	(0.040)
	[4.72]	[4.71]	[4.75]
Covariates that predict T	No	Yes	Yes
Covariates that predict Y	No	Yes	Yes
Covariates included	0	4	9

Standard errors in parentheses; t-statistics in brackets.

When Are Covariates Important?

$$MDE = \left(t_{1-\kappa} + t_{\alpha/2}\right) \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\tilde{\sigma}^2}{N}} \sqrt{1 + (n_g - 1)\,\tilde{\rho}}$$

Covariates can increase statistical power substantially when outcomes are serially correlated

- Covariates have limited impact when outcomes vary a lot over time
- Malaria example: residual variance 93 percent of original
- Including more covariates can increase power, but risks over-fitting