On Model Dependence in the Estimation of Interactive Effects

September 25th, 2019

Matthew Blackwell   Michael Olson

Motivation

effect heterogenetity

effect of treatment D_i is different at different levels of a moderator V_i

Motivation

effect heterogenetity

effect of treatment D_i is different at different levels of a moderator V_i

why do we care?

Motivation

effect heterogenetity

effect of treatment D_i is different at different levels of a moderator V_i

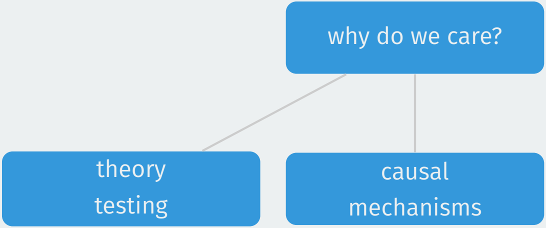
why do we care?

theory testing

Motivation

effect heterogenetity

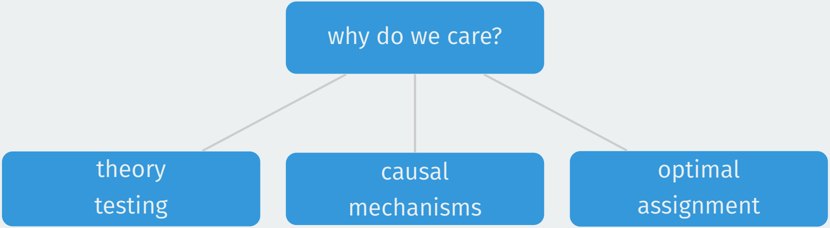
effect of treatment D_i is different at different levels of a moderator V_i



Motivation

effect heterogenetity

effect of treatment D_i is different at different levels of a moderator V_i



Two ways to investigate heterogeneity

split sample by
moderator

Two ways to investigate heterogeneity

split sample by moderator

single multiplicative interaction term

Two ways to investigate heterogeneity

split sample by moderator

Uncommon

single multiplicative interaction term

Two ways to investigate heterogeneity

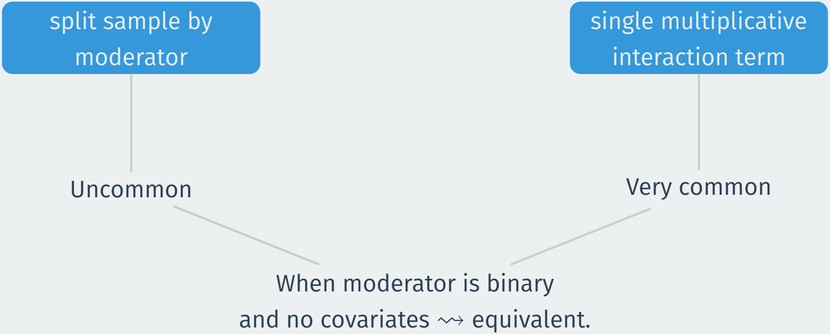
split sample by moderator

Uncommon

single multiplicative interaction term

Very common

Two ways to investigate heterogeneity



Two ways to investigate heterogeneity

split sample by moderator

Uncommon

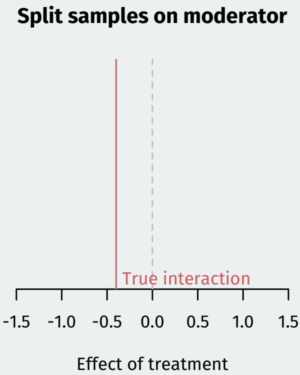
single multiplicative interaction term

Very common

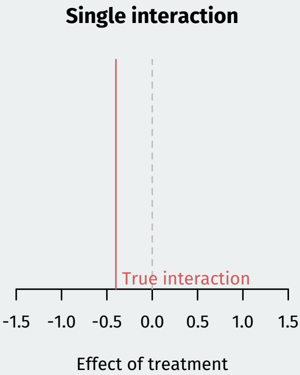
When moderator is binary
and no covariates $\rightsquigarrow$ equivalent.

...but can very different results in other conditions.

Toy Example

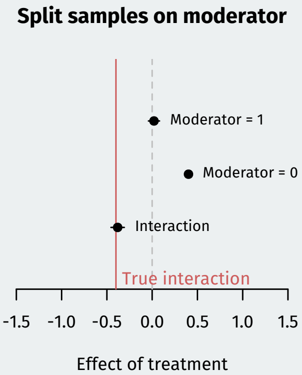


Split samples on moderator
True interaction
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Effect of treatment

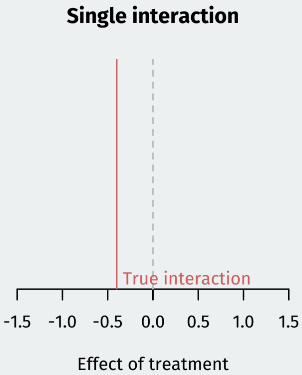


Single interaction
True interaction
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Effect of treatment

Toy Example

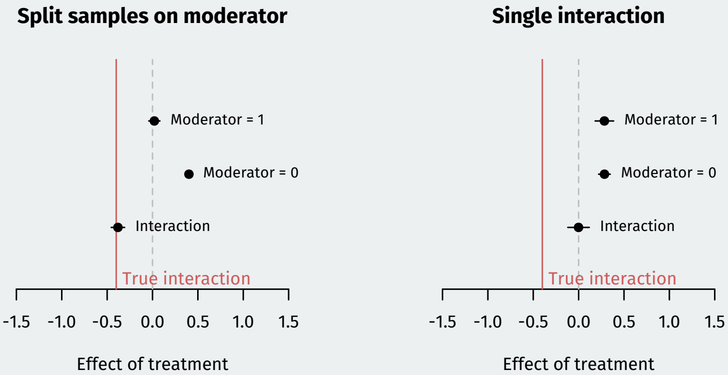


Split samples on moderator
Moderator = 1
Moderator = 0
Interaction
True interaction
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
Effect of treatment

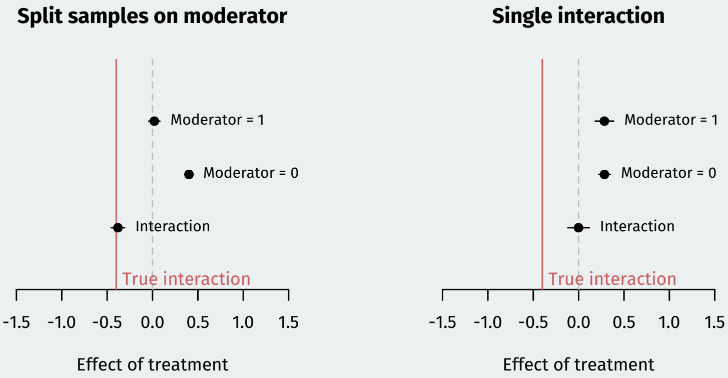


Single interaction
True interaction
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
Effect of treatment

Toy Example

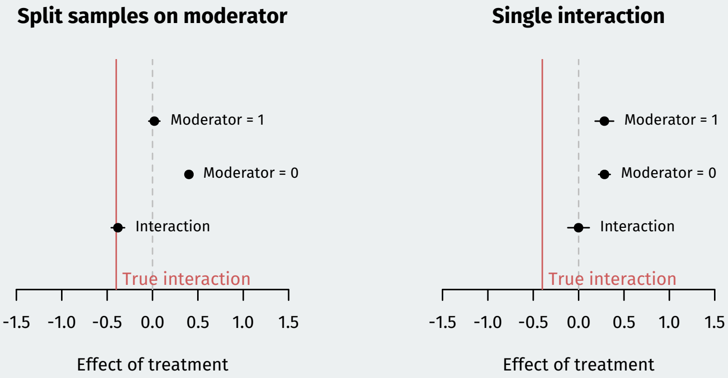

Split samples on moderator
Moderator = 1
Moderator = 0
Interaction
True interaction
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Effect of treatment
Single interaction
Moderator = 1
Moderator = 0
Interaction
True interaction
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Effect of treatment

Toy Example



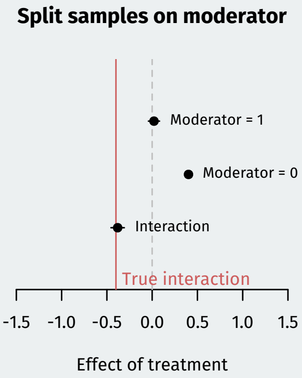
- Why do these approaches give different results?

Toy Example



- Why do these approaches give different results?
- Should we prefer one to the other?

Toy Example



- Why do these approaches give different results?
- Should we prefer one to the other?
- Is there another method that can outperform both?

Basic Problem

- Why the divergence? **Covariates!**

Basic Problem

- Why the divergence? **Covariates!**
  - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.

Basic Problem

- Why the divergence? **Covariates!**
    - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.
    - We call this **omitted interaction bias**, but really a form of model dependence.

Basic Problem

- Why the divergence? **Covariates!**
    - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.
    - We call this **omitted interaction bias**, but really a form of model dependence.
- Okay... so just use the split sample approach? (can we leave the talk early?)

Basic Problem

- Why the divergence? **Covariates!**
  - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.
  - We call this **omitted interaction bias**, but really a form of model dependence.
- Okay... so just use the split sample approach? (can we leave the talk early?)
  - Unfortunately, not always because... covariates!

Basic Problem

- Why the divergence? **Covariates!**
  - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.
  - We call this **omitted interaction bias**, but really a form of model dependence.
- Okay... so just use the split sample approach? (can we leave the talk early?)
  - Unfortunately, not always because... covariates!
  - Lots of covariates $\rightsquigarrow$ noisy estimates, overfitting.

Basic Problem

- Why the divergence? **Covariates!**
    - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.
    - We call this **omitted interaction bias**, but really a form of model dependence.
- Okay… so just use the split sample approach? (can we leave the talk early?)
    - Unfortunately, not always because… covariates!
    - Lots of covariates $\rightsquigarrow$ noisy estimates, overfitting.
- Our proposal: use **regularization** to balance between single interaction and split sample.

Basic Problem

- Why the divergence? **Covariates!**
  - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.
  - We call this **omitted interaction bias**, but really a form of model dependence.
- Okay... so just use the split sample approach? (can we leave the talk early?)
  - Unfortunately, not always because... covariates!
  - Lots of covariates $\rightsquigarrow$ noisy estimates, overfitting.
- Our proposal: use **regularization** to balance between single interaction and split sample.
  - Avoids overfitting while avoiding large biases of the single interaction.

Basic Problem

- Why the divergence? **Covariates!**
    - More specifically: a single-interaction model omits covariate-moderator interactions that are likely to be important.
    - We call this **omitted interaction bias**, but really a form of model dependence.
- Okay... so just use the split sample approach? (can we leave the talk early?)
    - Unfortunately, not always because... covariates!
    - Lots of covariates $\rightsquigarrow$ noisy estimates, overfitting.
- Our proposal: use **regularization** to balance between single interaction and split sample.
    - Avoids overfitting while avoiding large biases of the single interaction.
    - Can't just apply standard lasso due to bias, lack of uncertainty.

Interactions literature

- Cottage industry of interactions papers in political science covering:

Interactions literature

- Cottage industry of interactions papers in political science covering:
    - Finger wagging at omitting base terms, correct interpretation thereof.

Interactions literature

- Cottage industry of interactions papers in political science covering:
  - Finger wagging at omitting base terms, correct interpretation thereof.
  - Using plots to visualize marginal effects.

Interactions literature

- Cottage industry of interactions papers in political science covering:
  - Finger wagging at omitting base terms, correct interpretation thereof.
  - Using plots to visualize marginal effects.
  - Be careful of linearity assumptions with interactions.

Interactions literature

- Cottage industry of interactions papers in political science covering:
    - Finger wagging at omitting base terms, correct interpretation thereof.
    - Using plots to visualize marginal effects.
    - Be careful of linearity assumptions with interactions.
    - Do we need interactions in non-linear models?

Interactions literature

- Cottage industry of interactions papers in political science covering:
    - Finger wagging at omitting base terms, correct interpretation thereof.
    - Using plots to visualize marginal effects.
    - Be careful of linearity assumptions with interactions.
    - Do we need interactions in non-linear models?
- Statistics and causal inference literature focused on differences between “effect modfication” and “causal interaction.”

Interactions literature

- Cottage industry of interactions papers in political science covering:
    - Finger wagging at omitting base terms, correct interpretation thereof.
    - Using plots to visualize marginal effects.
    - Be careful of linearity assumptions with interactions.
    - Do we need interactions in non-linear models?
- Statistics and causal inference literature focused on differences between “effect modfication” and “causal interaction.”
- The issues here are orthogonal to most of this literature.

Interactions literature

- Cottage industry of interactions papers in political science covering:
    - Finger wagging at omitting base terms, correct interpretation thereof.
    - Using plots to visualize marginal effects.
    - Be careful of linearity assumptions with interactions.
    - Do we need interactions in non-linear models?
- Statistics and causal inference literature focused on differences between “effect modfication” and “causal interaction.”
- The issues here are orthogonal to most of this literature.
- Most similar: Vansteelandt, Vanderweele, Tchetgen Tchetgen, and Robins (2008, JASA) on multiply robust estimation of interactions.

Interactions literature

- Cottage industry of interactions papers in political science covering:
    - Finger wagging at omitting base terms, correct interpretation thereof.
    - Using plots to visualize marginal effects.
    - Be careful of linearity assumptions with interactions.
    - Do we need interactions in non-linear models?
- Statistics and causal inference literature focused on differences between “effect modfication” and “causal interaction.”
- The issues here are orthogonal to most of this literature.
- Most similar: Vansteelandt, Vanderweele, Tchetgen Tchetgen, and Robins (2008, JASA) on multiply robust estimation of interactions.
    - This approach still requires correct models somewhere, whereas we’ll use the lasso to select out the model.

Roadmap

1. The Problem
2. Solutions
3. Simulations
4. Empirical Applications
5. Conclusion

1/ The Problem

Setup and notation

- Assume iid sample $\{1, \dots, N\}$ (some clustering allowed later)

Setup and notation

- Assume iid sample $\{1, \dots, N\}$ (some clustering allowed later)
- Relevant variables:

Setup and notation

- Assume iid sample $\{1, \ldots, N\}$ (some clustering allowed later)
- Relevant variables:
  - Outcome Y_i, treatment D_i, and effect modifier V_i.

Setup and notation

- Assume iid sample $\{1, \ldots, N\}$ (some clustering allowed later)
- Relevant variables:
  - Outcome Y_i, treatment D_i, and effect modifier V_i.
  - Other pretreatment covariates: X_i of dimension K (might be high-dimensional)

Setup and notation

- Assume iid sample $\{1, \dots, N\}$ (some clustering allowed later)
- Relevant variables:
  - Outcome Y_i, treatment D_i, and effect modifier V_i.
  - Other pretreatment covariates: X_i of dimension K (might be high-dimensional)
- Important—we consider X_i to be **nuisances**.

Setup and notation

- Assume iid sample $\{1, \ldots, N\}$ (some clustering allowed later)
- Relevant variables:
  - Outcome Y_i, treatment D_i, and effect modifier V_i.
  - Other pretreatment covariates: X_i of dimension K (might be high-dimensional)
- Important—we consider X_i to be **nuisances**.
  - We only care about main effect of D_i and interaction with V_i.

Setup and notation

- Assume iid sample $\{1, \dots, N\}$ (some clustering allowed later)
- Relevant variables:
  - Outcome Y_i, treatment D_i, and effect modifier V_i.
  - Other pretreatment covariates: X_i of dimension K (might be high-dimensional)
- Important—we consider X_i to be **nuisances**.
  - We only care about main effect of D_i and interaction with V_i.
- Focusing on an **confirmatory** interaction analysis.

Setup and notation

- Assume iid sample $\{1, \dots, N\}$ (some clustering allowed later)
- Relevant variables:
  - Outcome Y_i, treatment D_i, and effect modifier V_i.
  - Other pretreatment covariates: X_i of dimension K (might be high-dimensional)
- Important—we consider X_i to be **nuisances**.
  - We only care about main effect of D_i and interaction with V_i.
- Focusing on an **confirmatory** interaction analysis.
  - Not directly interested in “exploring” all possible interactions between D_i and covariates.

Setup and notation

- Assume iid sample $\{1, \dots, N\}$ (some clustering allowed later)
- Relevant variables:
    - Outcome Y_i, treatment D_i, and effect modifier V_i.
    - Other pretreatment covariates: X_i of dimension K (might be high-dimensional)
- Important—we consider X_i to be **nuisances**.
    - We only care about main effect of D_i and interaction with V_i.
- Focusing on an **confirmatory** interaction analysis.
    - Not directly interested in “exploring” all possible interactions between D_i and covariates.
    - Dominant application of interactions in empirical papers.

Omitted interaction bias

Base regression $Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$

Omitted interaction bias

Base regression	$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$
Single interaction	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \boldsymbol{\beta}_3 + \beta_4 D_i V_i + \varepsilon_{i2}$

Omitted interaction bias

Base regression	$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$
Single interaction	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \boldsymbol{\beta}_3 + \beta_4 D_i V_i + \varepsilon_{i2}$
Fully moderated	$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$

Omitted interaction bias

Base regression	$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$
Single interaction	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \boldsymbol{\beta}_3 + \beta_4 D_i V_i + \varepsilon_{i2}$
Fully moderated	$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$
Omitted interaction bias	$\widehat{\beta}_4 \xrightarrow{p} \delta_4 + \boldsymbol{\gamma}_v' \boldsymbol{\delta}_5$

Omitted interaction bias

Base regression	$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$
Single interaction	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \boldsymbol{\beta}_3 + \beta_4 D_i V_i + \varepsilon_{i2}$
Fully moderated	$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$
Omitted interaction bias	$\widehat{\beta}_4 \xrightarrow{p} \delta_4 + \boldsymbol{\gamma}_v' \boldsymbol{\delta}_5$

- Single interaction assumes X_i have constant effects across V_i.

Omitted interaction bias

Base regression	$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$
Single interaction	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \boldsymbol{\beta}_3 + \beta_4 D_i V_i + \varepsilon_{i2}$
Fully moderated	$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$
Omitted interaction bias	$\widehat{\beta}_4 \xrightarrow{p} \delta_4 + \boldsymbol{\gamma}_v' \boldsymbol{\delta}_5$

- Single interaction assumes X_i have constant effects across V_i.
- Only valid when omitted interactions unrelated to Y_i ($\boldsymbol{\delta}_5 = 0$) or unrelated to $D_i V_i$ ($\boldsymbol{\gamma}_v = 0$).

Omitted interaction bias

Base regression	$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$
Single interaction	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \boldsymbol{\beta}_3 + \beta_4 D_i V_i + \varepsilon_{i2}$
Fully moderated	$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$
Omitted interaction bias	$\widehat{\beta}_4 \xrightarrow{p} \delta_4 + \boldsymbol{\gamma}_v' \boldsymbol{\delta}_5$

- Single interaction assumes X_i have constant effects across V_i.
- Only valid when omitted interactions unrelated to Y_i ($\boldsymbol{\delta}_5 = 0$) or unrelated to $D_i V_i$ ($\boldsymbol{\gamma}_v = 0$).
- Randomization of D_i does not guarantee that this holds.

Omitted interaction bias

Base regression	$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \boldsymbol{\alpha}_3 + \varepsilon_{i1}$
Single interaction	$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \boldsymbol{\beta}_3 + \beta_4 D_i V_i + \varepsilon_{i2}$
Fully moderated	$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$
Omitted interaction bias	$\widehat{\beta}_4 \xrightarrow{p} \delta_4 + \boldsymbol{\gamma}_v' \boldsymbol{\delta}_5$

- Single interaction assumes X_i have constant effects across V_i.
- Only valid when omitted interactions unrelated to Y_i ($\boldsymbol{\delta}_5 = 0$) or unrelated to $D_i V_i$ ($\boldsymbol{\gamma}_v = 0$).
- Randomization of D_i does not guarantee that this holds.
    - Holds if D_i and V_i are both randomized as in a conjoint experiment.

2/ Solutions

Easiest solutions

- Simplest solution: just run the fully moderated model.

Easiest solutions

- Simplest solution: just run the fully moderated model.
  - Avoids any omitted interaction bias.

Easiest solutions

- Simplest solution: just run the fully moderated model.
    - Avoids any omitted interaction bias.
    - Equivalent to splitting sample on V_i with easier uncertainty estimates.

Easiest solutions

- Simplest solution: just run the fully moderated model.
    - Avoids any omitted interaction bias.
    - Equivalent to splitting sample on V_i with easier uncertainty estimates.
- One could even generalize this to handle **heterogeneous effects**:

Easiest solutions

- Simplest solution: just run the fully moderated model.
    - Avoids any omitted interaction bias.
    - Equivalent to splitting sample on V_i with easier uncertainty estimates.
- One could even generalize this to handle **heterogeneous effects**:
    - Predicted values: $\hat{\mu}(d, v, x) = \widehat{\mathbb{E}}[Y_i \mid D_i = d, V_i = v, X_i = x]$

Easiest solutions

- Simplest solution: just run the fully moderated model.
  - Avoids any omitted interaction bias.
  - Equivalent to splitting sample on V_i with easier uncertainty estimates.
- One could even generalize this to handle **heterogeneous effects**:
  - Predicted values: $\hat{\mu}(d, v, x) = \widehat{\mathbb{E}}[Y_i \mid D_i = d, V_i = v, X_i = x]$
  - Estimated interaction:

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\mu}(1, 1, X_i) - \hat{\mu}(0, 1, X_i) - \hat{\mu}(1, 0, X_i) + \hat{\mu}(0, 0, X_i)$$

Easiest solutions

- Simplest solution: just run the fully moderated model.
    - Avoids any omitted interaction bias.
    - Equivalent to splitting sample on V_i with easier uncertainty estimates.
- One could even generalize this to handle **heterogeneous effects**:
    - Predicted values: $\hat{\mu}(d, v, x) = \widehat{\mathbb{E}}[Y_i \mid D_i = d, V_i = v, X_i = x]$
    - Estimated interaction:

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\mu}(1,1,X_i) - \hat{\mu}(0,1,X_i) - \hat{\mu}(1,0,X_i) + \hat{\mu}(0,0,X_i)$$

- Problem: if X_i is highly dimensional, fully moderated model will overfit and be noisy.

Easiest solutions

- Simplest solution: just run the fully moderated model.
  - Avoids any omitted interaction bias.
  - Equivalent to splitting sample on V_i with easier uncertainty estimates.
- One could even generalize this to handle **heterogeneous effects**:
  - Predicted values: $\hat{\mu}(d, v, x) = \widehat{\mathbb{E}}[Y_i \mid D_i = d, V_i = v, X_i = x]$
  - Estimated interaction:

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\mu}(1, 1, X_i) - \hat{\mu}(0, 1, X_i) - \hat{\mu}(1, 0, X_i) + \hat{\mu}(0, 0, X_i)$$

- Problem: if X_i is highly dimensional, fully moderated model will overfit and be noisy.
  - Roughly doubles the number of covariates in the model.

Easiest solutions

- Simplest solution: just run the fully moderated model.
  - Avoids any omitted interaction bias.
  - Equivalent to splitting sample on V_i with easier uncertainty estimates.
- One could even generalize this to handle **heterogeneous effects**:
  - Predicted values: $\hat{\mu}(d, v, x) = \widehat{\mathbb{E}}[Y_i \mid D_i = d, V_i = v, X_i = x]$
  - Estimated interaction:

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\mu}(1,1,X_i) - \hat{\mu}(0,1,X_i) - \hat{\mu}(1,0,X_i) + \hat{\mu}(0,0,X_i)$$

- Problem: if X_i is highly dimensional, fully moderated model will overfit and be noisy.
  - Roughly doubles the number of covariates in the model.
  - Can be substantial especially with fixed effects in X_i.

Regularization to the rescue?

- When free to pick any coefficients, OLS will pick very large values to minimize residuals $\rightsquigarrow$ overfitting.

Regularization to the rescue?

- When free to pick any coefficients, OLS will pick very large values to minimize residuals $\rightsquigarrow$ overfitting.
- Stabilize estimates via **regularization/shrinkage**: penalize coefficient vectors that are too large.

Regularization to the rescue?

- When free to pick any coefficients, OLS will pick very large values to minimize residuals $\rightsquigarrow$ overfitting.
- Stabilize estimates via **regularization/shrinkage**: penalize coefficient vectors that are too large.
- One popular approach: **Lasso** or L_1-regularization:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - X_i'\boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Regularization to the rescue?

- When free to pick any coefficients, OLS will pick very large values to minimize residuals $\rightsquigarrow$ overfitting.
- Stabilize estimates via **regularization/shrinkage**: penalize coefficient vectors that are too large.
- One popular approach: **Lasso** or L_1-regularization:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - X_i'\boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

  - $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$ is the L_1 norm of the coefficients.

Regularization to the rescue?

- When free to pick any coefficients, OLS will pick very large values to minimize residuals $\rightsquigarrow$ overfitting.
- Stabilize estimates via **regularization/shrinkage**: penalize coefficient vectors that are too large.
- One popular approach: **Lasso** or L_1-regularization:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - X_i'\boldsymbol{\beta}\right)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

  - $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$ is the L_1 norm of the coefficients.
  - $\lambda \geq 0$ is a **complexity parameter**: larger λ, more shrinkage.

Regularization to the rescue?

- When free to pick any coefficients, OLS will pick very large values to minimize residuals $\rightsquigarrow$ overfitting.
- Stabilize estimates via **regularization/shrinkage**: penalize coefficient vectors that are too large.
- One popular approach: **Lasso** or L_1-regularization:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - X_i'\boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

  - $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$ is the L_1 norm of the coefficients.
  - $\lambda \geq 0$ is a **complexity parameter**: larger λ, more shrinkage.
  - With large enough λ some coefficients will be set to 0 (sparsity).

Why the vanilla lasso doesn't work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \delta_1 D_i - \delta_2 V_i - X_i' \boldsymbol{\delta}_3 - \delta_4 D_i V_i - V_i X_i' \boldsymbol{\delta}_5\right)^2 + \lambda \|\boldsymbol{\delta}\|_1$$

Why the vanilla lasso doesn't work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \delta_1 D_i - \delta_2 V_i - X_i'\boldsymbol{\delta}_3 - \delta_4 D_i V_i - V_i X_i'\boldsymbol{\delta}_5\right)^2 + \lambda \|\boldsymbol{\delta}\|_1$$

Problems:

- Coefficients of interest are biased due to regularization, even in large samples.

Why the vanilla lasso doesn't work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \delta_1 D_i - \delta_2 V_i - X_i'\boldsymbol{\delta}_3 - \delta_4 D_i V_i - V_i X_i'\boldsymbol{\delta}_5\right)^2 + \lambda \|\boldsymbol{\delta}\|_1$$

Problems:

- Coefficients of interest are biased due to regularization, even in large samples.
- Bias due to costly model selection mistakes:

Why the vanilla lasso doesn't work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \delta_1 D_i - \delta_2 V_i - X_i'\boldsymbol{\delta}_3 - \delta_4 D_i V_i - V_i X_i'\boldsymbol{\delta}_5\right)^2 + \lambda \|\boldsymbol{\delta}\|_1$$

Problems:

- Coefficients of interest are biased due to regularization, even in large samples.
- Bias due to costly model selection mistakes:
    - Lasso will zero out interactions with small predictive power for Y_i, even if has massive importance for $D_i V_i$.

Why the vanilla lasso doesn't work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \delta_1 D_i - \delta_2 V_i - X_i'\boldsymbol{\delta}_3 - \delta_4 D_i V_i - V_i X_i'\boldsymbol{\delta}_5\right)^2 + \lambda ||\boldsymbol{\delta}||_1$$

Problems:

- Coefficients of interest are biased due to regularization, even in large samples.
- Bias due to costly model selection mistakes:
  - Lasso will zero out interactions with small predictive power for Y_i, even if has massive importance for $D_i V_i$.
- No straightforward way to obtain uncertainty estimates for QOIs.

Why the vanilla lasso doesn't work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \delta_1 D_i - \delta_2 V_i - X_i' \boldsymbol{\delta}_3 - \delta_4 D_i V_i - V_i X_i' \boldsymbol{\delta}_5\right)^2 + \lambda ||\boldsymbol{\delta}||_1$$

Problems:

- Coefficients of interest are biased due to regularization, even in large samples.
- Bias due to costly model selection mistakes:
  - Lasso will zero out interactions with small predictive power for Y_i, even if has massive importance for $D_i V_i$.
- No straightforward way to obtain uncertainty estimates for QOIs.
- Possible to select interaction while regularizing base term to 0 $\rightsquigarrow$ awkward interpretation.

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
    - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
    - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) “nuisance” variables.

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
  - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) “nuisance” variables.
- **Algorithm**:

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
    - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) “nuisance” variables.
- **Algorithm**:
    1. Run lasso of Y_i on Z_i with carefully chosen tuning parameter.

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
    - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) “nuisance” variables.
- **Algorithm**:
    1. Run lasso of Y_i on Z_i with carefully chosen tuning parameter.
    2. Run lasso of D_i on Z_i with carefully chosen tuning parameter.

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
    - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) "nuisance" variables.
- **Algorithm**:
    1. Run lasso of Y_i on Z_i with carefully chosen tuning parameter.
    2. Run lasso of D_i on Z_i with carefully chosen tuning parameter.
    3. Run lasso of $D_i V_i$ on Z_i with carefully chosen tuning parameter.

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
  - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) "nuisance" variables.
- **Algorithm**:
  1. Run lasso of Y_i on Z_i with carefully chosen tuning parameter.
  2. Run lasso of D_i on Z_i with carefully chosen tuning parameter.
  3. Run lasso of $D_i V_i$ on Z_i with carefully chosen tuning parameter.
  4. Collect variables selected (ie, non-zero) by any of (1)-(3) into Z_i^*

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
  - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) "nuisance" variables.
- **Algorithm**:
  1. Run lasso of Y_i on Z_i with carefully chosen tuning parameter.
  2. Run lasso of D_i on Z_i with carefully chosen tuning parameter.
  3. Run lasso of $D_i V_i$ on Z_i with carefully chosen tuning parameter.
  4. Collect variables selected (ie, non-zero) by any of (1)-(3) into Z_i^*
  5. Run OLS of Y_i on D_i, $D_i V_i$, and Z_i^*.

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
    - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) "nuisance" variables.
- **Algorithm**:
    1. Run lasso of Y_i on Z_i with carefully chosen tuning parameter.
    2. Run lasso of D_i on Z_i with carefully chosen tuning parameter.
    3. Run lasso of $D_i V_i$ on Z_i with carefully chosen tuning parameter.
    4. Collect variables selected (ie, non-zero) by any of (1)-(3) into Z_i^*
    5. Run OLS of Y_i on D_i, $D_i V_i$, and Z_i^*.
- One can optionally override the lasso for certain variables and force their inclusion into step (5).

Post-double selection procedure

- Our approach: adapt the **post-double-selection** approach of Belloni et al (2014) to our setting.
    - Originally designed to avoid regularization bias with high-dimensional covariates, but low dimensional quantities of interest (like the ATE).
- Let $Z_i' = [V_i, X_i', V_i X_i']$ be the vector of (centered) "nuisance" variables.
- **Algorithm**:
    1. Run lasso of Y_i on Z_i with carefully chosen tuning parameter.
    2. Run lasso of D_i on Z_i with carefully chosen tuning parameter.
    3. Run lasso of $D_i V_i$ on Z_i with carefully chosen tuning parameter.
    4. Collect variables selected (ie, non-zero) by any of (1)-(3) into Z_i^*
    5. Run OLS of Y_i on D_i, $D_i V_i$, and Z_i^*.
- One can optionally override the lasso for certain variables and force their inclusion into step (5).
    - We force all base terms to be included for comparison with other models.

Post-double-selection properties

- Avoids key biases:

Post-double-selection properties

- Avoids key biases:
    - Regularization bias avoided by post-lasso estimation via OLS.

Post-double-selection properties

- Avoids key biases:
    - Regularization bias avoided by post-lasso estimation via OLS.
    - Model selection mistakes avoided by taking union of variables important for outcome, treatment, and treatment-moderator interaction.

Post-double-selection properties

- Avoids key biases:
  - Regularization bias avoided by post-lasso estimation via OLS.
  - Model selection mistakes avoided by taking union of variables important for outcome, treatment, and treatment-moderator interaction.
- Belloni et al (2014) prove:

Post-double-selection properties

- Avoids key biases:
  - Regularization bias avoided by post-lasso estimation via OLS.
  - Model selection mistakes avoided by taking union of variables important for outcome, treatment, and treatment-moderator interaction.
- Belloni et al (2014) prove:
  - Coefficients on D_i and $D_i V_i$ are consistent.

Post-double-selection properties

- Avoids key biases:
  - Regularization bias avoided by post-lasso estimation via OLS.
  - Model selection mistakes avoided by taking union of variables important for outcome, treatment, and treatment-moderator interaction.
- Belloni et al (2014) prove:
  - Coefficients on D_i and $D_i V_i$ are consistent.
  - Standard errors from OLS asymptotically correct.

Post-double-selection properties

- Avoids key biases:
  - Regularization bias avoided by post-lasso estimation via OLS.
  - Model selection mistakes avoided by taking union of variables important for outcome, treatment, and treatment-moderator interaction.
- Belloni et al (2014) prove:
  - Coefficients on D_i and $D_i V_i$ are consistent.
  - Standard errors from OLS asymptotically correct.
  - Can allow for robust SEs as well.

Post-double-selection properties

- Avoids key biases:
  - Regularization bias avoided by post-lasso estimation via OLS.
  - Model selection mistakes avoided by taking union of variables important for outcome, treatment, and treatment-moderator interaction.
- Belloni et al (2014) prove:
  - Coefficients on D_i and $D_i V_i$ are consistent.
  - Standard errors from OLS asymptotically correct.
  - Can allow for robust SEs as well.
  - Can handle clustering as well, but requires different tuning parameter selection.

Approximate sparsity

- Belloni et al (2014) prove asymptotic results under key assumption of **approximate sparsity**:

$$\mathbb{E}[Y_i \mid Z_i] = Z_i'\boldsymbol{\delta}_{y0} + r_{yi},$$

$$\sum_{j=1}^{K} \mathbf{1}(\delta_{yj} \neq 0) \leq s, \qquad \left\{(1/N)\sum_i \mathbb{E}[r_{yi}^2]\right\}^{1/2} \leq C\sqrt{s/N}$$

  - CEFs are well-approximated by a sparse representation with s terms.
  - Similar assumptions on CEF for D_i and $D_i V_i$
  - Rate condition: $(s\log(\max(K,N)))^2/N \to 0$. Number of terms needed for approximation doesn't grow too quickly relative to N.
- Sample splitting can weaken this requirement, but difficult to apply with fixed effects which are common.

How to choose complexity parameter

$$\arg\min_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left(Y_i - Z_i' \boldsymbol{\delta}_y \right)^2 + \sum_{j=1}^{K} \lambda_{yj} |\delta_{yj}|$$

- Rate condition requires choosing penalty loadings carefully.

How to choose complexity parameter

$$\arg\min_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left(Y_i - Z_i'\boldsymbol{\delta}_y \right)^2 + \sum_{j=1}^{K} \lambda_{yj} |\delta_{yj}|$$

- Rate condition requires choosing penalty loadings carefully.
- Belloni et al show that the ideal penalty loadings for estimation (not prediction) are: $\lambda_{yj} \propto \sqrt{(1/N)\sum_i Z_{ij}^2 \varepsilon_i^2}$ where ε_i are the errors.

How to choose complexity parameter

$$\arg\min_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left(Y_i - Z_i' \boldsymbol{\delta}_y \right)^2 + \sum_{j=1}^{K} \lambda_{yj} |\delta_{yj}|$$

- Rate condition requires choosing penalty loadings carefully.
- Belloni et al show that the ideal penalty loadings for estimation (not prediction) are: $\lambda_{yj} \propto \sqrt{(1/N) \sum_i Z_{ij}^2 \varepsilon_i^2}$ where ε_i are the errors.
  - Intuition: more regularization for variables whose “noise” correlates with the error.

How to choose complexity parameter

$$\arg\min_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left(Y_i - Z_i' \boldsymbol{\delta}_y \right)^2 + \sum_{j=1}^{K} \lambda_{yj} |\delta_{yj}|$$

- Rate condition requires choosing penalty loadings carefully.
- Belloni et al show that the ideal penalty loadings for estimation (not prediction) are: $\lambda_{yj} \propto \sqrt{(1/N) \sum_i Z_{ij}^2 \varepsilon_i^2}$ where ε_i are the errors.
  - Intuition: more regularization for variables whose "noise" correlates with the error.
  - Feasible approach: run preliminary lasso to obtain estimates $\hat{\varepsilon}_i$.

How to choose complexity parameter

$$\arg\min_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left(Y_i - Z_i' \boldsymbol{\delta}_y \right)^2 + \sum_{j=1}^{K} \lambda_{yj} |\delta_{yj}|$$

- Rate condition requires choosing penalty loadings carefully.
- Belloni et al show that the ideal penalty loadings for estimation (not prediction) are: $\lambda_{yj} \propto \sqrt{(1/N) \sum_i Z_{ij}^2 \varepsilon_i^2}$ where ε_i are the errors.
  - Intuition: more regularization for variables whose “noise” correlates with the error.
  - Feasible approach: run preliminary lasso to obtain estimates $\hat{\varepsilon}_i$.
- Allows for non-normal and heteroskedastic errors.

How to choose complexity parameter

$$\arg\min_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left(Y_i - Z_i' \boldsymbol{\delta}_y \right)^2 + \sum_{j=1}^{K} \lambda_{yj} |\delta_{yj}|$$

- Rate condition requires choosing penalty loadings carefully.
- Belloni et al show that the ideal penalty loadings for estimation (not prediction) are: $\lambda_{yj} \propto \sqrt{(1/N) \sum_i Z_{ij}^2 \varepsilon_i^2}$ where ε_i are the errors.
    - Intuition: more regularization for variables whose "noise" correlates with the error.
    - Feasible approach: run preliminary lasso to obtain estimates $\hat{\varepsilon}_i$.
- Allows for non-normal and heteroskedastic errors.
- We apply an extension for clustered data in our applications (similar to cluster robust SEs).

3/ Simulations

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
  - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
  - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
  - Vary the number of covariates in X_i, $K \in \{20, 200\}$.

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
  - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
  - Vary the number of covariates in X_i, $K \in \{20, 200\}$.
- Note that this isn't a sparse model $\rightsquigarrow$ difficult case for lasso.

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i'\boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i'\boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i'\boldsymbol{\gamma}_2 + V_i X_i'\boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
  - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
  - Vary the number of covariates in X_i, $K \in \{20, 200\}$.
- Note that this isn't a sparse model $\rightsquigarrow$ difficult case for lasso.
- $N = 750$ and $10,000$ iterations per DGP.

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
  - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
  - Vary the number of covariates in X_i, $K \in \{20, 200\}$.
- Note that this isn't a sparse model $\rightsquigarrow$ difficult case for lasso.
- $N = 750$ and $10,000$ iterations per DGP.
- Methods to compare:

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
    - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
    - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
    - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
    - Vary the number of covariates in X_i, $K \in \{20, 200\}$.
- Note that this isn't a sparse model $\rightsquigarrow$ difficult case for lasso.
- $N = 750$ and $10,000$ iterations per DGP.
- Methods to compare:
    - Single interaction (not shown due to huge bias).

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
  - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
  - Vary the number of covariates in X_i, $K \in \{20, 200\}$.
- Note that this isn't a sparse model $\rightsquigarrow$ difficult case for lasso.
- $N = 750$ and $10,000$ iterations per DGP.
- Methods to compare:
  - Single interaction (not shown due to huge bias).
  - Fully moderated.

Simulation setup

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

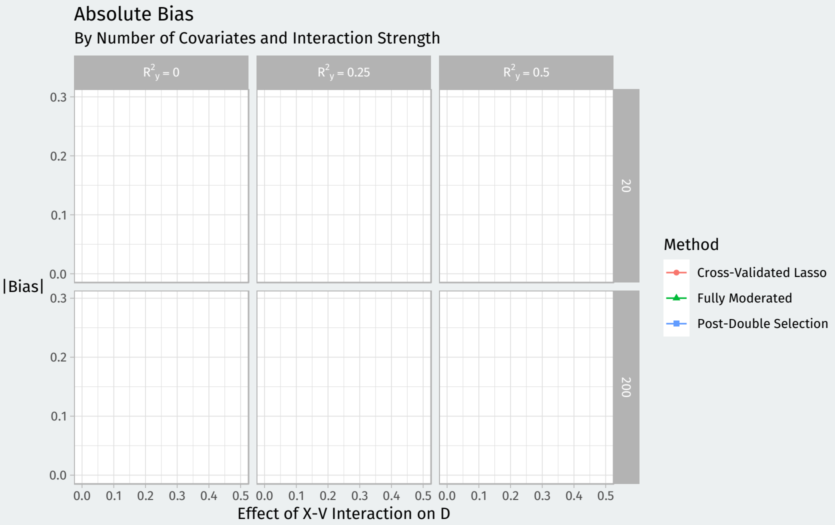
- DGP is fully moderated model where coefficients have quadratic decay:
  - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
  - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
  - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
  - Vary the number of covariates in X_i, $K \in \{20, 200\}$.
- Note that this isn't a sparse model $\rightsquigarrow$ difficult case for lasso.
- $N = 750$ and $10,000$ iterations per DGP.
- Methods to compare:
  - Single interaction (not shown due to huge bias).
  - Fully moderated.
  - Post-lasso on just outcome (using cross-validation).

Simulation setup

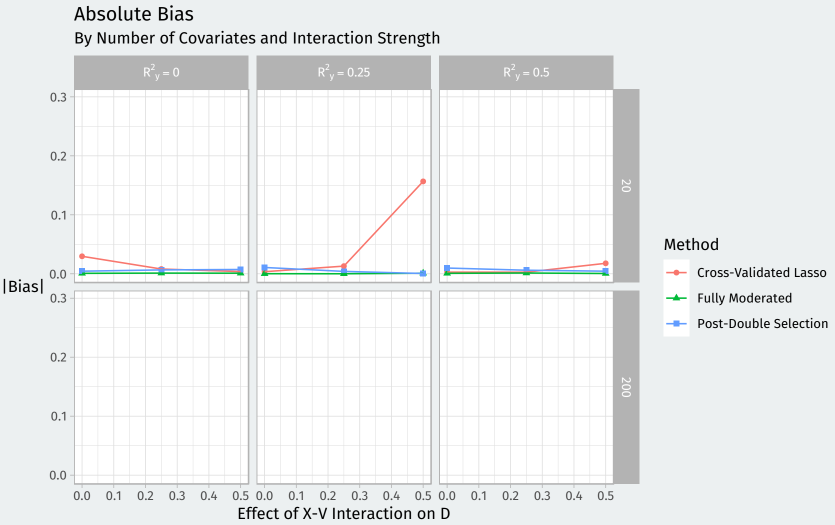
$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \boldsymbol{\delta}_3 + \delta_4 D_i V_i + V_i X_i' \boldsymbol{\delta}_5 + \varepsilon_{i3}$$
$$D_i = \gamma_0 + \gamma_1 V_i + X_i' \boldsymbol{\gamma}_2 + V_i X_i' \boldsymbol{\gamma}_3$$

- DGP is fully moderated model where coefficients have quadratic decay:
    - Effect of X-V interactions on Y: $\delta_{5j} = c_{vy}(1/j^2)$
    - Effect of X-V interactions on D: $\gamma_{3j} = c_{vd}(1/j^2)$
    - Select c_{vy} and c_{vd} to have partial R^2 of these interaction terms be in $\{0, 0.25, 0.5\}$.
    - Vary the number of covariates in X_i, $K \in \{20, 200\}$.
- Note that this isn't a sparse model $\rightsquigarrow$ difficult case for lasso.
- $N = 750$ and $10,000$ iterations per DGP.
- Methods to compare:
    - Single interaction (not shown due to huge bias).
    - Fully moderated.
    - Post-lasso on just outcome (using cross-validation).
    - Post-double-selection.

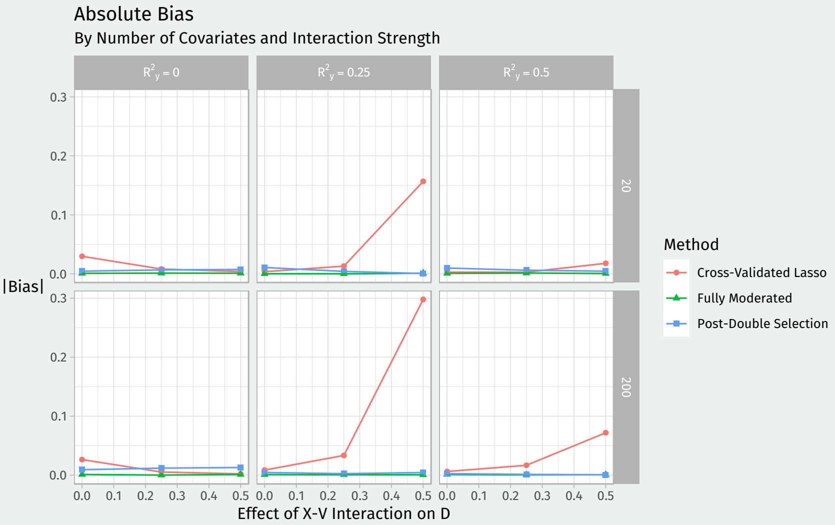
Simulation results: bias



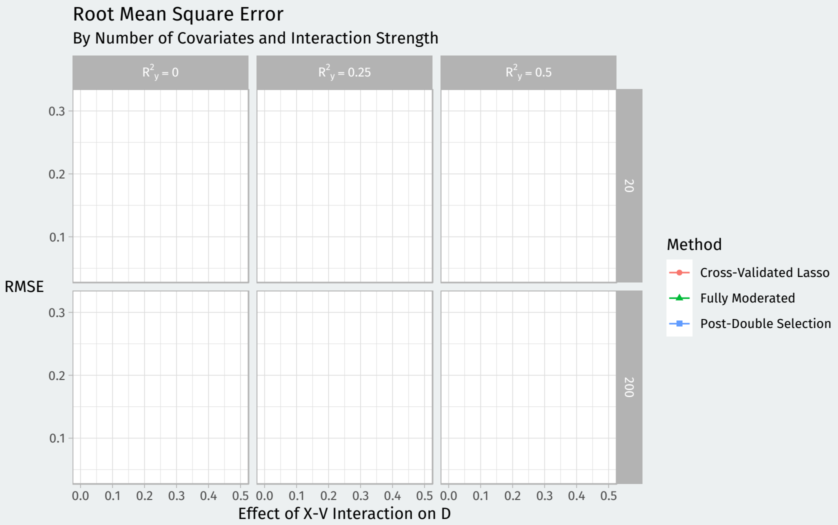
Simulation results: bias



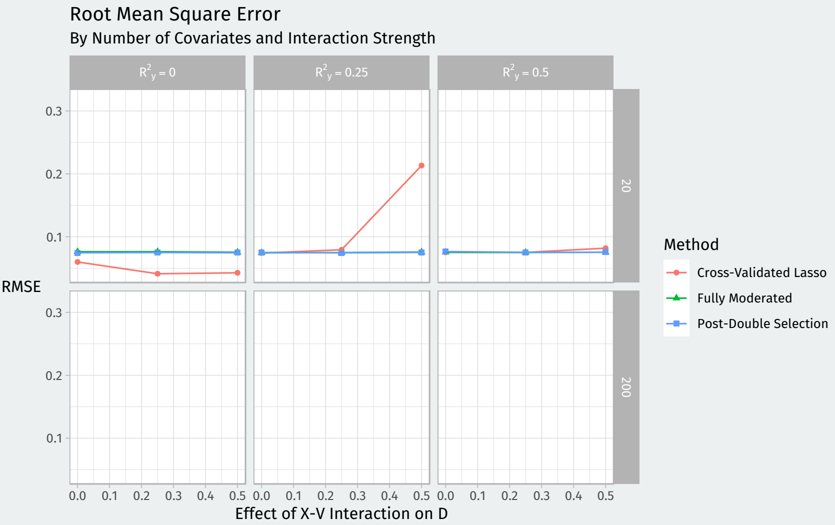
Simulation results: bias



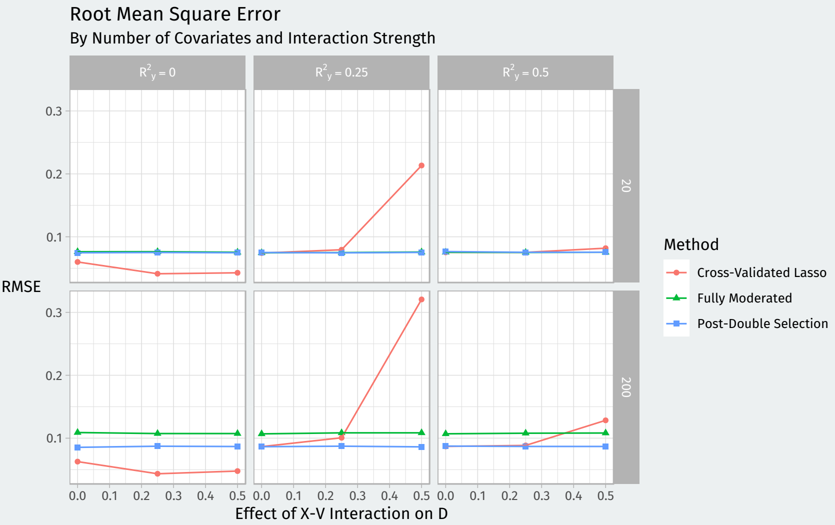
Simulation results: RMSE



Simulation results: RMSE



Simulation results: RMSE



4/ Empirical Applications

Regime type and remittances

- Escribà-Folch, Meseguer, and Wright (AJPS 2018) argue that higher levels of incoming remittances ought to lead to higher levels of political protest, but only in autocracies

Regime type and remittances

- Escribà-Folch, Meseguer, and Wright (AJPS 2018) argue that higher levels of incoming remittances ought to lead to higher levels of political protest, but only in autocracies
- “We show that remittances are associated with protests in autocratic regimes, but not in democracies.” (890)

Regime type and remittances

- Escribà-Folch, Meseguer, and Wright (AJPS 2018) argue that higher levels of incoming remittances ought to lead to higher levels of political protest, but only in autocracies
- "We show that remittances are associated with protests in autocratic regimes, but not in democracies." (890)
- Pair novel continuous measure of protest (based on dynamic IRT) with World Development Indicators data on remittances remittances entering a country

Regime type and remittances

- Escribà-Folch, Meseguer, and Wright (AJPS 2018) argue that higher levels of incoming remittances ought to lead to higher levels of political protest, but only in autocracies
- "We show that remittances are associated with protests in autocratic regimes, but not in democracies." (890)
- Pair novel continuous measure of protest (based on dynamic IRT) with World Development Indicators data on remittances remittances entering a country
- 102 non-OECD countries (coded as democracies or autocracies) from 1976 to 2010

Regime type and remittances

Original Model

$$Protest_{it} = \beta\left(Remit_{it} \times Autocracy_{it}\right) + \gamma Remit_{it} + \phi Autocracy_{it} + \boldsymbol{\psi}\mathbf{X}_{it} + \tau_t + \alpha_i + \epsilon_{it}$$

- Quantity of interest is β: coefficient on single interaction between remittances (continuous) and autocracies (binary)

Regime type and remittances

Original Model

$$Protest_{it} = \beta\,(Remit_{it} \times Autocracy_{it}) + \gamma\,Remit_{it} + \phi Autocracy_{it} + \boldsymbol{\psi}\mathbf{X}_{it} + \tau_t + \alpha_i + \epsilon_{it}$$

- Quantity of interest is β: coefficient on single interaction between remittances (continuous) and autocracies (binary)
- Model includes country (α) and five-year time period (τ) fixed effects

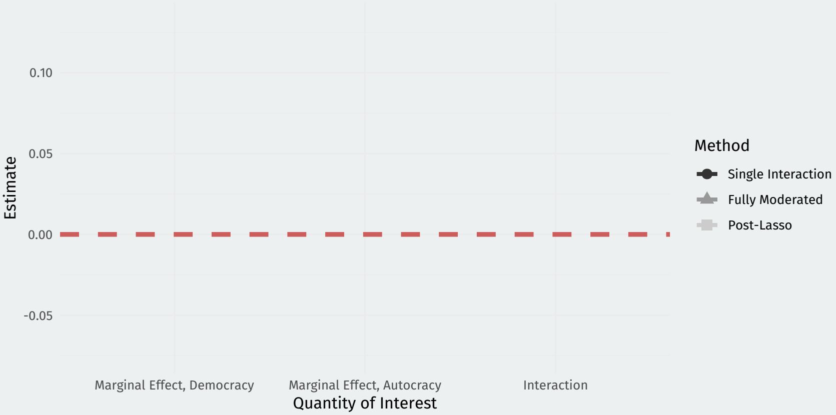
Regime type and remittances

Original Model

$$Protest_{it} = \beta\left(Remit_{it} \times Autocracy_{it}\right) + \gamma Remit_{it} + \phi Autocracy_{it} + \boldsymbol{\psi}\mathbf{X}_{it} + \tau_t + \alpha_i + \epsilon_{it}$$

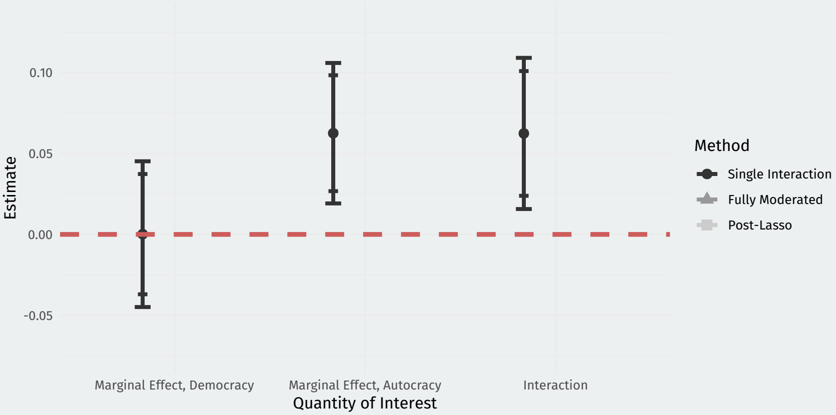
- Quantity of interest is β: coefficient on single interaction between remittances (continuous) and autocracies (binary)
- Model includes country (α) and five-year time period (τ) fixed effects
- $\mathbf{X}$ is a vector of time-varying covariates

Results

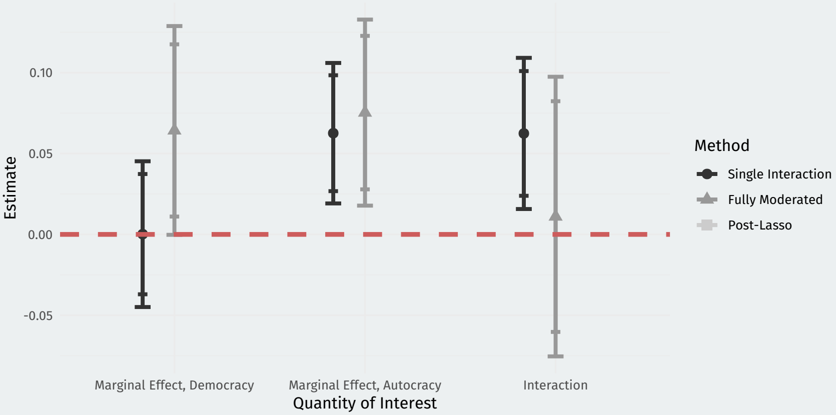


0.10
0.05
0.00
-0.05
Estimate
Marginal Effect, Democracy
Marginal Effect, Autocracy
Interaction
Quantity of Interest
Method
Single Interaction
Fully Moderated
Post-Lasso

Results

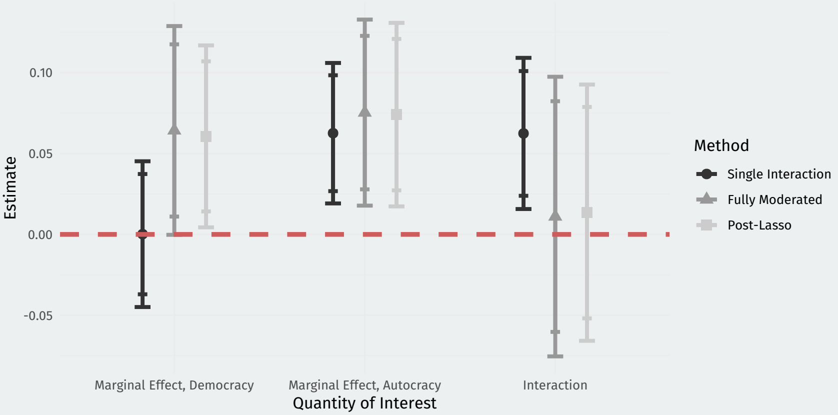

0.10
0.05
0.00
-0.05
Estimate
Marginal Effect, Democracy
Marginal Effect, Autocracy
Interaction
Quantity of Interest
Method
Single Interaction
Fully Moderated
Post-Lasso

Results



Estimate
0.10
0.05
0.00
-0.05
Marginal Effect, Democracy
Marginal Effect, Autocracy
Interaction
Quantity of Interest
Method
Single Interaction
Fully Moderated
Post-Lasso

Results



Estimate
0.10
0.05
0.00
-0.05
Marginal Effect, Democracy
Marginal Effect, Autocracy
Interaction
Quantity of Interest
Method
Single Interaction
Fully Moderated
Post-Lasso

5/ Conclusion

Summary

- When estimating interactions, interactions on “nuisance” covariates can be important.

Summary

- When estimating interactions, interactions on “nuisance” covariates can be important.
- Single interaction model $\rightsquigarrow$ omitted interaction bias.

Summary

- When estimating interactions, interactions on "nuisance" covariates can be important.
- Single interaction model $\rightsquigarrow$ omitted interaction bias.
- Fully moderated models (split sample on moderator) can avoid these bias.

Summary

- When estimating interactions, interactions on “nuisance” covariates can be important.
- Single interaction model $\rightsquigarrow$ omitted interaction bias.
- Fully moderated models (split sample on moderator) can avoid these bias.
- We propose an alternative when dimensionality of covariates is high: post-double-selection using the lasso.

Summary

- When estimating interactions, interactions on “nuisance” covariates can be important.
- Single interaction model $\rightsquigarrow$ omitted interaction bias.
- Fully moderated models (split sample on moderator) can avoid these bias.
- We propose an alternative when dimensionality of covariates is high: post-double-selection using the lasso.
    - Performs well against alternatives even in finite samples.

Summary

- When estimating interactions, interactions on “nuisance” covariates can be important.
- Single interaction model $\rightsquigarrow$ omitted interaction bias.
- Fully moderated models (split sample on moderator) can avoid these bias.
- We propose an alternative when dimensionality of covariates is high: post-double-selection using the lasso.
    - Performs well against alternatives even in finite samples.
    - Post-double-selection more broadly useful for estimating treatment effects with high-dimensional covariates.

Summary

- When estimating interactions, interactions on “nuisance” covariates can be important.
- Single interaction model $\rightsquigarrow$ omitted interaction bias.
- Fully moderated models (split sample on moderator) can avoid these bias.
- We propose an alternative when dimensionality of covariates is high: post-double-selection using the lasso.
    - Performs well against alternatives even in finite samples.
    - Post-double-selection more broadly useful for estimating treatment effects with high-dimensional covariates.
- Next steps:

Summary

- When estimating interactions, interactions on “nuisance” covariates can be important.
- Single interaction model $\rightsquigarrow$ omitted interaction bias.
- Fully moderated models (split sample on moderator) can avoid these bias.
- We propose an alternative when dimensionality of covariates is high: post-double-selection using the lasso.
    - Performs well against alternatives even in finite samples.
    - Post-double-selection more broadly useful for estimating treatment effects with high-dimensional covariates.
- Next steps:
    - Apply the split-sample approach of the double machine learning literature to this setting to relax some assumptions.

Thanks!

For more information...

Matt Blackwell

mattblackwell.org

@matt_blackwell

Michael Olson

michaelpatrickolson.com

@michael_p_olson