On Model Dependence in the Estimation of Interactive Effects

September 25th, 2019
Matthew Blackwell    Michael Olson
Effect heterogeneity

Effect of treatment \( D_i \) is different at different levels of a moderator \( V_i \)
Motivation

effect heterogeneity

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

why do we care?
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 heterogeneity

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

- why do we care?

- theory testing

- causal mechanisms
Motivation

- effect heterogeneity

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

why do we care?

- theory testing
- causal mechanisms
- optimal assignment
Two ways to investigate heterogeneity

- split sample by moderator
- single multiplicative interaction term

Uncommon

Very common

When moderator is binary ⇝ equivalent.

…but can very different results in other conditions.
Two ways to investigate heterogeneity

split sample by moderator

single multiplicative interaction term
Two ways to investigate heterogeneity

- split sample by moderator
- single multiplicative interaction term

Uncommon

When moderator is binary and no covariates ⇝ equivalent.

…but can very different results in other conditions.
Two ways to investigate heterogeneity

- split sample by moderator
  - Uncommon
- single multiplicative interaction term
  - Very common

When moderator is binary and no covariates ⇝ equivalent.

…but can very different results in other conditions.
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.
Two ways to investigate heterogeneity

- **split sample by moderator**
  - Uncommon
  - When moderator is binary and no covariates $\rightarrow$ equivalent.
  - ...but can very different results in other conditions.

- **single multiplicative interaction term**
  - Very common
Toy Example

Split samples on moderator

<table>
<thead>
<tr>
<th>-1.5</th>
<th>-1.0</th>
<th>-0.5</th>
<th>0.0</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
</tr>
</thead>
</table>

Single interaction

<table>
<thead>
<tr>
<th>-1.5</th>
<th>-1.0</th>
<th>-0.5</th>
<th>0.0</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
</tr>
</thead>
</table>

Effect of treatment

True interaction

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

Split samples on moderator

Single interaction

-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5

Moderator = 1

Moderator = 0

Interaction

True interaction

Effect of treatment

Effect of treatment

Why do these approaches give different results?

Should we prefer one to the other?

Is there another method that can outperform both?
Toy Example

- **Split samples on moderator**
  - Moderator = 1
  - Moderator = 0
  - Interaction
  - True interaction
  - Effect of treatment

- **Single interaction**
  - Moderator = 1
  - Moderator = 0
  - Interaction
  - True interaction
  - Effect of treatment

Why do these approaches give different results?
Should we prefer one to the other?
Is there another method that can outperform both?
Why do these approaches give different results?
• Why do these approaches give different results?
• Should we prefer one to the other?
Why do these approaches give different results?
Should we prefer one to the other?
Is there another method that can outperform both?
• 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.
• 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?)
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!
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 $\leadsto$ 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 ⇒ 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 $\Rightarrow$ 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 $\Rightarrow$ 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:

  - 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 modification” and “causal interaction.”
  - The issues here are orthogonal to most of this literature.

  - This approach still requires correct models somewhere, whereas we’ll use the lasso to select out the model.
• Cottage industry of interactions papers in political science covering:
  • Finger wagging at omitting base terms, correct interpretation thereof.
• 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.
• 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 modification” 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 modification” 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 modification” 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 modification” and “causal interaction.”

• The issues here are orthogonal to most of this literature.

  • This approach still requires correct models somewhere, whereas we’ll use the lasso to select out the model.
1. The Problem

2. Solutions

3. Simulations

4. Empirical Applications

5. Conclusion
1/ The Problem
• Assume iid sample \( \{1, \ldots, N\} \) (some clustering allowed later)
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 \(V_i\).
  - Focusing on a confirmatory interaction analysis.
  - Not directly interested in “exploring” all possible interactions between \(D_i\) and covariates.
  - Dominant application of interactions in empirical papers.
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, \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.
• 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 \( V_i \).
• 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 \(V_i\).
• Focusing on an **confirmatory** interaction analysis.
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 \(V_i\).
• 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, \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 \(V_i\).
• 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

\[ Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \epsilon_{i1} \]

- Single interaction assumes \( X_i \) have constant effects across \( V_i \).
- Only valid when omitted interactions unrelated to \( Y_i \) (\( \delta_5 = 0 \)) or unrelated to \( D_i V_i \) (\( \delta_4 = 0 \)).
- 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.
Omitted interaction bias

Base regression

\[ Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \varepsilon_{i1} \]

Single interaction

\[ Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \varepsilon_{i2} \]
Omitted interaction bias

**Base regression**

\[ Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \epsilon_{i1} \]

**Single interaction**

\[ Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \epsilon_{i2} \]

**Fully moderated**

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \epsilon_{i3} \]
Omitted interaction bias

Base regression

\[ Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \varepsilon_{i1} \]

Single interaction

\[ Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \varepsilon_{i2} \]

Fully moderated

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \]

\[ \hat{\beta}_4 \overset{p}{\rightarrow} \delta_4 + \gamma'_V \delta_5 \]
Omitted interaction bias

- **Base regression**
  \[ Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \varepsilon_{i1} \]

- **Single interaction**
  \[ Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \varepsilon_{i2} \]

- **Fully moderated**
  \[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \]

- **Omitted interaction bias**
  \[ \hat{\beta}_4 \overset{p}{\rightarrow} \delta_4 + \gamma'_V \delta_5 \]

- Single interaction assumes \( X_i \) have constant effects across \( V_i \).
Omitted interaction bias

**Base regression**

\[ Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \epsilon_{i1} \]

**Single interaction**

\[ Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \epsilon_{i2} \]

**Fully moderated**

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \epsilon_{i3} \]

\[ \hat{\beta}_4 \overset{p}{\rightarrow} \delta_4 + y'_v \delta_5 \]

- Single interaction assumes \( X_i \) have constant effects across \( V_i \).
- Only valid when omitted interactions unrelated to \( Y_i \) (\( \delta_5 = 0 \)) or unrelated to \( D_i V_i \) (\( y'_v = 0 \)).
## Omitted interaction bias

<table>
<thead>
<tr>
<th>Type</th>
<th>Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base regression</strong></td>
<td>$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i'\alpha_3 + \varepsilon_{i1}$</td>
</tr>
<tr>
<td><strong>Single interaction</strong></td>
<td>$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i'\beta_3 + \beta_4 D_i V_i + \varepsilon_{i2}$</td>
</tr>
<tr>
<td><strong>Fully moderated</strong></td>
<td>$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i'\delta_3 + \delta_4 D_i V_i + V_i X_i'\delta_5 + \varepsilon_{i3}$</td>
</tr>
</tbody>
</table>

### Omitted interaction bias

- $\beta_4 \xrightarrow{p} \delta_4 + \gamma_V'\delta_5$

- Single interaction assumes $X_i$ have constant effects across $V_i$.
- Only valid when omitted interactions unrelated to $Y_i$ ($\delta_5 = 0$) or unrelated to $D_i V_i$ ($\gamma_V = 0$).
- Randomization of $D_i$ does not guarantee that this holds.
Omitted interaction bias

**Base regression**

\[ Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \epsilon_{i1} \]

**Single interaction**

\[ Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \epsilon_{i2} \]

**Fully moderated**

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \epsilon_{i3} \]

\[ \hat{\beta}_4 \xrightarrow{p} \delta_4 + \gamma'_v \delta_5 \]

- Single interaction assumes \( X_i \) have constant effects across \( V_i \).
- Only valid when omitted interactions unrelated to \( Y_i \) (\( \delta_5 = 0 \)) or unrelated to \( D_i V_i \) (\( \gamma'_v = 0 \)).
- 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.
Solutions
• Simplest solution: just run the fully moderated model.
Easiest solutions

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

\[
\hat{\mu}(d, v, x) = \hat{E}[Y_i | D_i = d, V_i = v, X_i = x]
\]

\[
\text{Estimated interaction: } \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}(1, 1, X_i) - \hat{\mu}(0, 1, X_i) - \hat{\mu}(1, 0, X_i) + \hat{\mu}(0, 0, X_i)
\]

• 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 \( X_i \).
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.

- 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 $X_i$. 
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: $\hat{\mu}(d, v, x) = \hat{E}[Y_i \mid D_i = d, V_i = v, X_i = x]$
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: $\hat{\mu}(d, v, x) = \hat{\mathbb{E}}[Y_i \mid D_i = d, V_i = v, X_i = x]$
  • 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: $\hat{\mu}(d, v, x) = \hat{E}[Y_i \mid D_i = d, V_i = v, X_i = x]$
  - Estimated interaction:

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

- 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: $\hat{\mu}(d, v, x) = \hat{\mathbb{E}}[Y_i \mid D_i = d, V_i = v, X_i = x]$  
  - 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: $\hat{\mu}(d, v, x) = \hat{E}[Y_i \mid D_i = d, V_i = v, X_i = x]$ 
  • 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 $\rightarrow$ overfitting.

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

- $\|\beta\|_1 = \sum_j |\beta_j|$ is the $L_1$ norm of the coefficients.
- $\lambda \geq 0$ is a complexity parameter: larger $\lambda$, more shrinkage.
- With large enough $\lambda$ some coefficients will be set to 0 (sparsity).
• When free to pick any coefficients, OLS will pick very large values to minimize residuals ⟷ 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 $\Rightarrow$ overfitting.
- Stabilize estimates via **regularization/shrinkage**: penalize coefficient vectors that are too large.
- One popular approach: **Lasso** or $L_1$-regularization:

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

- $\|\beta\|_1 = \sum_j |\beta_j|$ is the $L_1$ norm of the coefficients.
- $\lambda \geq 0$ is a complexity parameter: larger $\lambda$, more shrinkage.
- With large enough $\lambda$ some coefficients will be set to 0 (sparsity).
• When free to pick any coefficients, OLS will pick very large values to minimize residuals \(\iff\) overfitting.

• Stabilize estimates via **regularization/shrinkage**: penalize coefficient vectors that are too large.

• One popular approach: **Lasso** or \(L_1\)-regularization:

\[
\hat{\beta}_{\text{lasso}} = \arg \min_\beta \sum_{i=1}^{N} (Y_i - X_i' \beta)^2 + \lambda \|\beta\|_1
\]

• \(\|\beta\|_1 = \sum_j |\beta_j|\) is the \(L_1\) norm of the coefficients.
• When free to pick any coefficients, OLS will pick very large values to minimize residuals $\implies$ overfitting.

• Stabilize estimates via regularization/shrinkage: penalize coefficient vectors that are too large.

• One popular approach: Lasso or $L_1$-regularization:

\[
\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1
\]

• $\|\beta\|_1 = \sum_j |\beta_j|$ is the $L_1$ norm of the coefficients.

• $\lambda \geq 0$ is a complexity parameter: larger $\lambda$, more shrinkage.
When free to pick any coefficients, OLS will pick very large values to minimize residuals \(\rightarrow\) overfitting.

Stabilize estimates via regularization/shrinkage: penalize coefficient vectors that are too large.

One popular approach: Lasso or \(L_1\)-regularization:

\[
\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1
\]

\(\|\beta\|_1 = \sum_j |\beta_j|\) is the \(L_1\) norm of the coefficients.

\(\lambda \geq 0\) is a complexity parameter: larger \(\lambda\), more shrinkage.

With large enough \(\lambda\) some coefficients will be set to 0 (sparsity).
Why the vanilla lasso doesn’t work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \delta_1 D_i - \delta_2 V_i - X_i' \delta_3 - \delta_4 D_i V_i - V_i X_i' \delta_5)^2 + \lambda \| \delta \|_1$$
Why the vanilla lasso doesn’t work

One solution: Apply standard lasso to fully moderated model:

$$\arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \delta_1 D_i - \delta_2 V_i - X_i' \delta_3 - \delta_4 D_i V_i - V_i X_i' \delta_5)^2 + \lambda \| \delta \|_1$$

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:

\[
\arg \min_{\beta} \sum_{i=1}^{N} \left( Y_i - \delta_1 D_i - \delta_2 V_i - X_i' \delta_3 - \delta_4 D_i V_i - V_i X_i' \delta_5 \right)^2 + \lambda \| \delta \|_1
\]

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:

$$\arg \min_{\beta} \sum_{i=1}^{N} \left( Y_i - \delta_1 D_i - \delta_2 V_i - X_i' \delta_3 - \delta_4 D_i V_i - V_i X_i' \delta_5 \right)^2 + \lambda \| \delta \|_1$$

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:

\[
\arg \min_\beta \sum_{i=1}^{N} \left( Y_i - \delta_1 D_i - \delta_2 V_i - X'_i \delta_3 - \delta_4 D_i V_i - V_i X'_i \delta_5 \right)^2 + \lambda \| \delta \|_1
\]

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:

\[
\arg \min_\beta \sum_{i=1}^{N} (Y_i - \delta_1 D_i - \delta_2 V_i - X'_i \delta_3 - \delta_4 D_i V_i - V_i X'_i \delta_5)^2 + \lambda \| \delta \|_1
\]

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 \( \leadsto \) awkward interpretation.
Post-double selection procedure

- Our approach: adapt the post-double-selection approach of Belloni et al (2014) to our setting.
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 $Z_i' = [V_i, X_i', V_iX_i']$ 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 $Z_i' = [V_i, X_i', V_iX_i']$ 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 $D_iV_i$ 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 $D_i$, $D_iV_i$, 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 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 \( Z_i' = [V_i, X_i', V_iX_i'] \) be the vector of (centered) “nuisance” variables.
• Algorithm:
  1. Run lasso of \( Y_i \) on \( Z_i \) with carefully chosen tuning parameter.
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 \( Z_i' = [V_i, X_i', V_iX_i'] \) 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.

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 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.

• 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.
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 $Z'_i = [V_i, X'_i, V_iX'_i]$ 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 $D_iV_i$ on $Z_i$ with carefully chosen tuning parameter.
4. Collect variables selected (ie, non-zero) by any of (1)-(3) into $Z_i^*$
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 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 \( Z_i' = [V_i, X_i', V_iX_i'] \) 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 \( D_iV_i \) 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 \( D_i, D_iV_i, \) and \( Z_i^* \).

- One can optionally override the lasso for certain variables and force their inclusion into step (5).
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 $Z_i' = [V_i, X_i', V_iX_i']$ 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 $D_iV_i$ 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 $D_i, D_iV_i,$ 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:
  - 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.
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.

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.
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 $D_iV_i$ 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 $D_iV_i$ 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 $D_iV_i$ 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:

\[ \mathbb{E}[Y_i \mid Z_i] = Z_i' \delta_{y0} + r_{yi}, \]

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

• 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

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

- Rate condition requires choosing penalty loadings carefully.

Rate condition requires choosing penalty loadings carefully.

Belloni et al show that the ideal penalty loadings for estimation (not prediction) are:

\[ \lambda_{y_j} \propto \sqrt{\frac{1}{N} \sum_{i} Z_{i}^2 \epsilon_i^2} \]
where \( \epsilon_i \) are the errors.

Intuition: more regularization for variables whose “noise” correlates with the error.

Feasible approach: run preliminary lasso to obtain estimates \( \hat{\epsilon}_i \).

Allows for non-normal and heteroskedastic errors.

We apply an extension for clustered data in our applications (similar to cluster robust SEs).
How to choose complexity parameter

\[
\arg \min_{\delta} \sum_{i=1}^{N} \left( Y_i - Z_i' \delta_y \right)^2 + \sum_{j=1}^{K} \lambda_y |\delta_{yj}| 
\]

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

\[
\arg \min_{\delta} \sum_{i=1}^{N} \left( Y_i - Z_i' \delta_y \right)^2 + \sum_{j=1}^{K} \lambda_{y_j} |\delta_{y_j}|
\]

- Rate condition requires choosing penalty loadings carefully.
- Belloni et al show that the ideal penalty loadings for estimation (not prediction) are: \( \lambda_{y_j} \propto \sqrt{(1/N) \sum_i Z_{i,j}^2 \epsilon_i^2} \) where \( \epsilon_i \) are the errors.
  - Intuition: more regularization for variables whose “noise” correlates with the error.
How to choose complexity parameter

\[
\arg \min_{\delta} \sum_{i=1}^{N} (Y_i - Z_i' \delta_y)^2 + \sum_{j=1}^{K} \lambda_{y_j} |\delta_{y_j}|
\]

- Rate condition requires choosing penalty loadings carefully.
- Belloni et al show that the ideal penalty loadings for estimation (not prediction) are: \( \lambda_{y_j} \propto \sqrt{(1/N) \sum_i Z_{ij}^2 \epsilon_i^2} \) where \( \epsilon_i \) are the errors.
  - Intuition: more regularization for variables whose “noise” correlates with the error.
  - Feasible approach: run preliminary lasso to obtain estimates \( \hat{\epsilon}_i \).
How to choose complexity parameter

\[ \arg \min_{\delta} \sum_{i=1}^{N} \left( Y_i - Z'_i \delta_y \right)^2 + \sum_{j=1}^{K} \lambda_{y_j} |\delta_{y_j}| \]

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

\[
\arg\min_{\delta} \sum_{i=1}^{N} \left( Y_i - Z_i' \delta_y \right)^2 + \sum_{j=1}^{K} \lambda_{y,j} |\delta_{y,j}| 
\]

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

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- DGP is fully moderated model where coefficients have quadratic decay:
\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \epsilon_{i3} \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

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

• Note that this isn’t a sparse model \( \Rightarrow \) difficult case for lasso.

• 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 setup

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- 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 \( D \): \( \gamma_{3j} = c_{vd}(1/j^2) \)
\begin{align*}
Y_i &= \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i \delta_3 + \delta_4 D_i V_i + V_i X_i \delta_5 + \epsilon_i^3 \\
D_i &= \gamma_0 + \gamma_1 V_i + X_i \gamma_2 + V_i X_i \gamma_3
\end{align*}

- 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

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_i \]

\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- 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 \( X_i, K \in \{20, 200\} \).
\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X'_i \delta_3 + \delta_4 D_i V_i + V_i X'_i \delta_5 + \varepsilon_{i3} \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X'_i \gamma_2 + V_i X'_i \gamma_3 \]

- 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 \( \sim \) difficult case for lasso.
Simulation setup

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \]

\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- 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 \( \leadsto \) difficult case for lasso.
- \( N = 750 \) and 10,000 iterations per DGP.
\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \epsilon_{i3} \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- 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 \( \sim \) difficult case for lasso.
- \( N = 750 \) and 10,000 iterations per DGP.
- Methods to compare:
Simulation setup

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_i \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- 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 \( \Rightarrow \) 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

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_i \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- 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

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

- 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 \( \sim \) 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

\[ Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \]
\[ D_i = \gamma_0 + \gamma_1 V_i + X_i' \gamma_2 + V_i X_i' \gamma_3 \]

• 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 \( \Rightarrow \) 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

Absolute Bias
By Number of Covariates and Interaction Strength

Effect of X-V Interaction on D

<table>
<thead>
<tr>
<th>Bias</th>
<th>R²_y = 0</th>
<th>R²_y = 0.25</th>
<th>R²_y = 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Method
- Cross-Validated Lasso
- Fully Moderated
- Post-Double Selection
Simulation results: bias

Absolute Bias
By Number of Covariates and Interaction Strength

<table>
<thead>
<tr>
<th>Effect of X-V Interaction on D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2_y = 0$</td>
</tr>
<tr>
<td>$R^2_y = 0.25$</td>
</tr>
<tr>
<td>$R^2_y = 0.5$</td>
</tr>
</tbody>
</table>

Method
- Cross-Validated Lasso
- Fully Moderated
- Post-Double Selection
Simulation results: bias

Absolute Bias
By Number of Covariates and Interaction Strength

<table>
<thead>
<tr>
<th>R²_y = 0</th>
<th>R²_y = 0.25</th>
<th>R²_y = 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Method
- Cross-Validated Lasso
- Fully Moderated
- Post-Double Selection
Simulation results: RMSE

Root Mean Square Error
By Number of Covariates and Interaction Strength

Effect of X-V Interaction on D

Method
- Cross-Validated Lasso
- Fully Moderated
- Post-Double Selection
Simulation results: RMSE

Root Mean Square Error
By Number of Covariates and Interaction Strength

Method
- Cross-Validated Lasso
- Fully Moderated
- Post-Double Selection
Simulation results: RMSE

Root Mean Square Error
By Number of Covariates and Interaction Strength

- $R^2_y = 0$
- $R^2_y = 0.25$
- $R^2_y = 0.5$

Effect of X-V Interaction on D
Empirical Applications
• 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
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)
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 entering a country.
• 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

\[
\text{Protest}_{it} = \beta (\text{Remit}_{it} \times \text{Autocracy}_{it}) + \gamma \text{Remit}_{it} + \phi \text{Autocracy}_{it} + \psi \mathbf{x}_{it} + \tau_t + \alpha_i + \epsilon_{it}
\]

- Quantity of interest is $\beta$: coefficient on single interaction between remittances (continuous) and autocracies (binary)
Regime type and remittances

Original Model

\[
Protest_{it} = \beta \left( \text{Remit}_{it} \times \text{Autocracy}_{it} \right) + \gamma \text{Remit}_{it} \\
+ \phi \text{Autocracy}_{it} + \psi X_{it} + \tau_t + \alpha_i + \epsilon_{it}
\]

- Quantity of interest is $\beta$: coefficient on single interaction between remittances (continuous) and autocracies (binary)
- Model includes country ($\alpha$) and five-year time period ($\tau$) fixed effects
Regime type and remittances

Original Model

$$Protest_{it} = \beta (Remit_{it} \times Autocracy_{it}) + \gamma Remit_{it} + \phi Autocracy_{it} + \psi X_{it} + \tau_t + \alpha_i + \epsilon_{it}$$

- Quantity of interest is $\beta$: coefficient on single interaction between remittances (continuous) and autocracies (binary)
- Model includes country ($\alpha$) and five-year time period ($\tau$) fixed effects
- $X$ is a vector of time-varying covariates
Results

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>Single Interaction</td>
</tr>
<tr>
<td>0.00</td>
<td>Fully Moderated</td>
</tr>
<tr>
<td>0.00</td>
<td>Post-Lasso</td>
</tr>
</tbody>
</table>

Marginal Effect, Democracy  
Marginal Effect, Autocracy  
Interaction  

Quantity of Interest
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Marginal Effect, Democracy</th>
<th>Marginal Effect, Autocracy</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Method</td>
<td>Single Interaction</td>
<td>Fully Moderated</td>
<td>Post-Lasso</td>
</tr>
</tbody>
</table>

The graph shows the estimates for Marginal Effect, Democracy and Marginal Effect, Autocracy, along with their interaction effect, across different methods: Single Interaction, Fully Moderated, and Post-Lasso. The estimates are depicted with error bars indicating the uncertainty around each estimate. The x-axis represents the Quantity of Interest, while the y-axis shows the Estimate range from -0.05 to 0.10.
Results

<table>
<thead>
<tr>
<th>Quantity of Interest</th>
<th>Estimate</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Single Interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fully Moderated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Post-Lasso</td>
</tr>
</tbody>
</table>

Marginal Effect, Democracy

Marginal Effect, Autocracy

Interaction

Method
- Single Interaction
- Fully Moderated
- Post-Lasso
Results

-0.05
0.00
0.05
0.10
Marginal Effect, Democracy Marginal Effect, Autocracy Interaction Quantity of Interest Estimate Method Single Interaction Fully Moderated Post-Lasso
Conclusion
• When estimating interactions, interactions on “nuisance” covariates can be important.
• When estimating interactions, interactions on “nuisance” covariates can be important.

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

• Single interaction model $\leadsto$ 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.
• When estimating interactions, interactions on “nuisance” covariates can be important.

• Single interaction model $\leadsto$ 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.
• When estimating interactions, interactions on “nuisance” covariates can be important.

• Single interaction model $\leadsto$ 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.
• When estimating interactions, interactions on “nuisance” covariates can be important.

• Single interaction model $\Rightarrow$ 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:
• When estimating interactions, interactions on “nuisance” covariates can be important.
• Single interaction model $\implies$ 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