Post-selection inference for effect modification

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

September 4, 2024 @ RSS Conference, Brighton

Acknowledgement

- Qingyuan Zhao, Dylan S. Small, and Ashkan Ertefaie (2022). "Selective Inference for Effect Modification via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84.2, pp. 382–413. DOI: 10.1111/rssb.12483 (arXiv: 1705.08020).
- Qingyuan Zhao and Snigdha Panigrahi (2019). "Selective Inference for Effect Modification: An Empirical Investigation". In: *Observational Studies* 5.2, pp. 131–140. DOI: 10.1353/obs.2019.0007.
- In hindsight, should be **post-selection inference** instead of **selective inference**.

Effect modification

- A treatment has different effects in different subgroups.
- Central problem in precision medicine/data integration/understanding causal mechanism.
- Common approach: subgroup or regression analysis with treatment-covariate interactions.

Cental problem

Can we make valid inference after *post hoc* selection of subgroups/interactions?

Example

In a special workshop of the 2018 Atlantic Causal Inference Conferenc, the organizers provided a dataset simulated from the National Study of Learning Mindsets and posed three questions:

1. Is the intervention effective in improving student achievement?
2. Do two hypothesized covariates (`X1` and `X2`) moderate the treatment effect?
3. Are there other covariates moderating the treatment effect?

Problem setup

- We observe i.i.d. variables $(\boldsymbol{X}_i, T_i, Y_i)$, $i = 1, \ldots, n$.
- We assume a non-parametric model for potential outcomes:

$$Y_i(t) = \eta(\boldsymbol{X}_i) + t \cdot \Delta(\boldsymbol{X}_i) + \epsilon_i(t), \; i = 1, \ldots, n,$$

  where $E\{\epsilon_i(t) \mid \boldsymbol{X}_i\} = 0$.
- When T_i is binary, this model is saturated and $\Delta(\boldsymbol{x}) = E\{Y_i(1) - Y_i(0) \mid X_i = x\}$ is the *conditional average treatment effect*.
- We make the usual causal identification assumptions: consistency/SUTVA, unconfoundedness, positivity/overlap.

Trading off between model accuracy and interpretability

	Univariate model	**Selected submodel**	Full linear model	Machine learning
Model of $\Delta(\boldsymbol{X}_i)$	$\alpha_j + X_{ij}\beta_j$	$\alpha_{\hat{\mathcal{M}}} + \boldsymbol{X}_{i,\hat{\mathcal{M}}}^T \boldsymbol{\beta}_{\hat{\mathcal{M}}}$	$\alpha + \boldsymbol{X}_i^T \boldsymbol{\beta}$	e.g. additive trees
Accuracy	Poor	**Good**	Good	Very good
Interpretability	Very good	**Good**	Poor	Very poor
Inference	Easy, but many false positives	**Need to consider model selection**	Semiparametric theory	Not clear **(new: conformal)**

Our solution

1. Use the transformation in Robinson (1988) to eliminate the nuisance parameter $\eta(\boldsymbol{X}_i)$:
$$Y_i - \mu_y(\boldsymbol{X}_i) = \{T_i - \mu_t(\boldsymbol{X}_i)\} \cdot \Delta(\boldsymbol{X}_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$
where $\mu_y(\boldsymbol{X}_i) = E(Y_i \mid \boldsymbol{X}_i)$ and $\mu_t(\boldsymbol{X}_i) = E(T_i \mid \boldsymbol{X}_i)$.
2. Select an effect modification model by solving
$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \underset{\alpha, \boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} \left[\{Y_i - \hat{\mu}_y(\boldsymbol{X}_i)\} - \{T_i - \hat{\mu}_t(\boldsymbol{X}_i)\} \cdot (\alpha - \mathbf{X}_i^T \boldsymbol{\beta})\right]^2 + \lambda \|\boldsymbol{\beta}\|_1$$
and letting $\hat{\mathcal{M}} = \{j : \hat{\beta}_j \neq 0\}$.
3. Use the pivotal statistic in Lee, Sun, Sun, and Taylor (2016) to obtain post-selection confidence intervals for the projection parameters
$$\boldsymbol{\beta}^*_{\hat{\mathcal{M}}} = \boldsymbol{\beta}^*_{\hat{\mathcal{M}}}(\mathbf{T}, \mathbf{X}) = \underset{\alpha, \boldsymbol{\beta}_{\hat{\mathcal{M}}}}{\arg\min} \sum_{i=1}^{n} \{T_i - \mu_t(\mathbf{X}_i)\}^2 \{\Delta(\mathbf{X}_i) - \alpha - \mathbf{X}_{i,\hat{\mathcal{M}}}^T \boldsymbol{\beta}_{\hat{\mathcal{M}}}\}^2.$$

- Implementation is straightforward using off-the-shelf machine learning packages (to estimate μ_y and μ_t) and existing software for post-selection inference.

Some theory

Theorem

We assume:

1. $\boldsymbol{X}_i$ *has bounded support.*
2. *Rate conditions in Robinson (1988):* $\|\hat{\mu}_t - \mu_t\|_2 = o_p(n^{-1/4})$, $\|\hat{\mu}_y - \mu_y\|_2 = o_p(1)$, $\|\hat{\mu}_t - \mu_t\|_2 \cdot \|\hat{\mu}_y - \mu_y\|_2 = o_p(n^{-1/2})$.
3. *Size of the selected model is bounded.*
4. *Lasso selected model is "stable" and not a small probability event.*

We then show that the post-selection confidence intervals are asymptotically valid **given the selected model**.

Back to the Mindset Study

1. Is the intervention effective in improving student achievement?

   Solution Use the no interaction model $\Delta(\boldsymbol{x}) = \alpha$.

2. Do two hypothesized covariates (`X1` and `X2`) moderate the treatment effect?

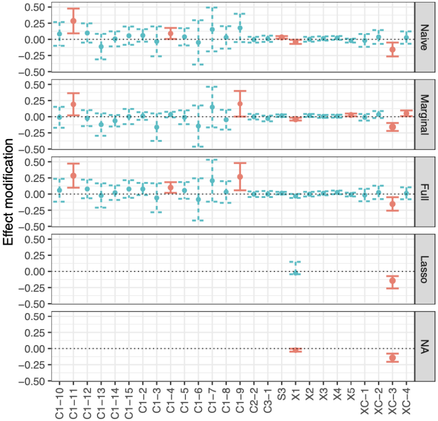
   Solution Use the pre-specified interaction model $\Delta(\boldsymbol{x}) = \alpha + x_1\beta_1 + x_2\beta_2$.

3. Are there other covariates moderating the treatment effect?

   Solution Use the post-hoc interaction model $\Delta(\boldsymbol{x}) = \alpha + \boldsymbol{x}_{\hat{\mathcal{M}}}^T \beta_{\hat{\mathcal{M}}}$.

Results

- (Weighted) average treatment effect is 0.256, with 95% CI [0.235, 0.277].



Methods compared:

- **Naive:** Linear model with all treatment-covariate interactions.
- **Marginal:** After Robinson's transformation, fits univariable regressions.
- **Full:** After Robinson's transformation, fits a full linear regression.
- **Lasso:** The proposed method.
- **Snooping** (incorrectly labeled as **NA**): Ignore model selection.

- Conclusion: `X1` is an effect modifier, `X2` is not, and using the data we discovered another effect modifier `XC-3`.

Discussion and reflection after 7 years

- The proposed method achieves a good trade-off between accuracy and interpretability. In particular, the final model for effect modification is familiar to applied statisticians.
- However, there are many caveats:
  1. The whole method is based on Robinson's transformation for partially linear models and post-selection inference for linear models.
  2. Inference is made for some (weighted) projection parameters.
  3. Some assumptions in the asymptotic theory look strong.
  4. A sufficient adjustment set (to control for confounding) is assumed to be given.
- Possible future directions:
  1. Genearlize the methodology/theory using semiparametric and post-selection inference.
  2. Data-adaptive confounder selection and post-selection inference.
  3. Sensitivity analysis.