Selective Inference for Effect Modification

Qingyuan Zhao
(Joint work with Dylan Small and Ashkan Ertefaie)

Department of Statistics, University of Pennsylvania

May 24, ACIC 2017

Manuscript and slides are available at
http://www-stat.wharton.upenn.edu/~qyzhao/.
Effect modification

Effect modification means the treatment has a different effect among different subgroups.

In other words, there is interaction between treatment and covariates in the outcome model.

Why care about effect modification?
- Extrapolation of average causal effect to a different population.
- Personalizing the treatment.
- Understanding the causal mechanism.
Subgroup analysis and regression analysis are the most common ways to analyze effect modification.

- Prespecified subgroups/interactions:
  - Free of selection bias. Scientifically rigorous.
  - Limited in number. No flexibility.

- Post hoc subgroups/interactions.
  - Scheffé, Tukey (1950s): multiple comparisons.
  - Lots of recent work on discovering effect modification.
  - But how to guarantee coverage? A call for valid inference after model selection.
Setting

A nonparametric model for the potential outcomes:

\[ Y_i(t) = \eta(X_i) + t \cdot \Delta(X_i) + \epsilon_i(t), \quad i = 1, \ldots, n. \]

- \( \Delta(x) \) is the parameter of interest.
- Saturated if the treatment is binary, \( t \in \{0, 1\} \).

Basic assumptions:

<table>
<thead>
<tr>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>
Naive linear modeling I

A straw man

Instead of the nonparametric model,

\[ Y_i(t) = \eta(X_i) + t \cdot \Delta(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \]

fit a linear model (the intercepts are dropped for simplicity)

\[ Y_i(t) = \gamma^T X_i + T_i \cdot (\beta^T X_i) + \tilde{\epsilon}_i, \quad i = 1, \ldots, n. \]

Dismiss all insignificant interaction terms, then refit the model.
Two critical fallacies:

1. Linear model could be misspecified.
   - **Solution: use machine learning algorithms to estimate the nuisance parameters.**
   - Targeted learning [van der Laan and Rose, 2011], double machine learning [Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, et al., 2016].

2. Statistical inference ignored data snooping.
   - **Solution: use selective inference.**
   - Lee, Sun, Sun, and Taylor [2016], Fithian, Sun, and Taylor [2014], Tian and Taylor [2017b].
Background: valid inference after model selection

- Acknowledge that the model is selected using the data.
- Model selection procedure:
  \[ \{X_i, T_i, Y_i\}_{i=1}^n \mapsto \hat{M} \quad (\text{data} \mapsto \text{a subset of covariates}) \]
- The target parameter \( \beta^*_\hat{M} \) is defined by \( \hat{M} \): \( x_{\hat{M}}^T \beta^*_\hat{M} \) is the “best linear approximation” of \( \Delta(x) \) [Berk, Brown, Buja, Zhang, and Zhao, 2013].
Background: valid inference after model selection II

Two types of confidence intervals:

1. **Simultaneous coverage** [Berk et al., 2013]:

   \[
   P\left( (\beta^*_\hat{M})_j \in [D^-_j, D^+_j] \text{ for any } j \in \hat{M} \right) \geq 1 - q, \ \forall \hat{M}.
   \]

2. **Conditional coverage** [Lee et al., 2016]:

   \[
   P\left( (\beta^*_M)_j \in [D^-_j, D^+_j] \bigg| \hat{M} = M \right) \geq 1 - q, \ \forall M.
   \]

Guarantees the control of false coverage rate (FCR, the average proportion of non-covering intervals among the reported) [Benjamini and Yekutieli, 2005].
Suppose we have noisy observations of $\Delta$:

$$Y_i = \Delta(X_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

- Submodel parameter
  $$\beta^*_M = \arg\min_{\alpha, \beta_M} \sum_{i=1}^{n} \left( \Delta(X_i) - \alpha - X_{i,M}^T \beta_M \right)^2.$$

- Linear selection rule
  $$\{ \hat{M} = M \} = \{ A_M(X) \cdot Y \leq b_M(X) \}.$$

- Example: Nonzero elements in the Lasso solution.
Background: selective inference in linear models II

- Main result of Lee et al. [2016]:
  \[
  (\hat{\beta}_{\hat{M}})_j \mid AY \leq b \text{ is truncated normal with mean } (\beta^*_{\hat{M}})_j.
  \]

- Need normality of noise, but can be relaxed in large sample [Tian and Taylor, 2017a].

- Geometric intuition:

  - Invert the pivotal statistic \( F((\hat{\beta}_{\hat{M}})_j, (\beta^*_{\hat{M}})_j) \sim \text{Unif}(0, 1) \) to construct selective confidence interval.
Eliminate the nuisance parameter

- Back to the causal model (of the observables)

\[ Y_i = \eta(X_i) + T_i \cdot \Delta(X_i) + \epsilon_i, \quad i = 1, \ldots, n. \]

- Problem: how to eliminate the nuisance parameter \( \eta(x) \)?

**Robinson [1988]'s transformation**

Let \( \mu_y(x) = \mathbb{E}[Y_i|X_i = x] \) and \( \mu_t(x) = \mathbb{E}[T_i|X_i = x] \), so

\[ \mu_y(x) = \eta(x) + \mu_t(x)\Delta(x). \]

An equivalent model is

\[ Y_i - \mu_y(X_i) = (T_i - \mu_t(X_i)) \cdot \Delta(X_i) + \epsilon_i, \quad i = 1, \ldots, n. \]

- The new nuisance parameters \( \mu_y(x) \) and \( \mu_t(x) \) can be directly estimated from the data.
Our complete proposal

- Estimate $\mu_y(x)$ and $\mu_t(x)$ using machine learning algorithms (for example random forest).

- Select a model for effect modification by solving

$$
\min_{\alpha, \beta} \sum_{i=1}^{n} \left[ (Y_i - \hat{\mu}_y(X_i)) - (T_i - \hat{\mu}_t(X_i)) \cdot (\alpha + X_i^T \beta) \right]^2 + \lambda \| \beta \|_1.
$$

- Use the pivotal statistic in Lee et al. [2016] to obtain selective confidence intervals of

$$
\beta^*_\hat{M} = \arg \min_{\alpha, \beta, \hat{M}} \sum_{i=1}^{n} (T_i - \mu_t(X_i))^2 (\Delta(X_i) - \alpha - X_{i,\hat{M}}^T \beta_{\hat{M}})^2.
$$
Main result

Challenge: $\mu_y$ and $\mu_t$ are estimated (with error).

Assumption

Rate assumptions in Robinson [1988]:

\[
\|\hat{\mu}_t - \mu_t\|_\infty = o_p\left(n^{-1/4}\right), \quad \|\hat{\mu}_y - \mu_y\|_\infty = o_p(1),
\]

\[
\|\hat{\mu}_t - \mu_t\|_\infty \cdot \|\hat{\mu}_y - \mu_y\|_\infty = o_p\left(n^{-1/2}\right).
\]

Theorem

Under additional assumptions on the selection event, the pivotal statistic and hence the selective confidence interval is asymptotically valid.
Simulation

Idealized estimation error

The true design and the true outcome were generated by

\[ X_i \sim \mathcal{N}(0, I), \quad Y_i \sim \mathcal{N}(X_i^T \beta, 1), \quad i = 1, \ldots, n, \]

where \( \beta = (1, 1, 1, 0, \ldots, 0)^T \in \mathbb{R}^{30}. \)

Then the design and the outcome were perturbed by

\[ X_i \mapsto X_i \cdot (1 + n^{-\gamma} e_{1i}), \quad Y_i \mapsto Y_i + n^{-\gamma} e_{2i}, \]

where \( e_{1i} \) and \( e_{2i} \) are independent standard normal.

- In the paper we also evaluated the validity of the entire procedure.
Rate assumptions are necessary and sufficient

- Crucial rate assumption: \( \|\hat{\mu}_t - \mu_t\|_\infty \cdot \|\hat{\mu}_y - \mu_y\|_\infty = o_p(n^{-1/2}) \).
- Phase transition at \( \gamma = 0.25 \):
  - When \( \gamma > 0.25 \): FCR is controlled at 10%.
  - When \( \gamma < 0.25 \): FCR is not controlled.
Real data example I


- Obesity was linked with systemic inflammation in the body. Prespecified subgroup analysis found effect modification by gender. Within women, they found effect modification by age group.

- We used a more recent dataset from NHANES 2007–2008 and 2009–2010.
Real data example II

- \( T \): obesity (BMI \( \geq 25 \)).
- \( Y \): C-reactive protein level.
- \( X \): gender, age, income, race, marital status, education, vigorous work activity (yes or no), vigorous recreation activities (yes or no), ever smoked, number of cigarettes smoked in the last month, estrogen usage, and if the survey respondent had bronchitis, asthma, emphysema, thyroid, arthritis, heart attack, stroke, liver condition, gout, and all their interactions.
- \( n = 9677, p = 355 \).
- \( \mu_y(x) \) and \( \mu_t(x) \) are estimated by \texttt{randomForest} in \texttt{R}.
- By running our procedure, lasso found two effect modifiers: gender and age (no surprise!).
### Real data example III

<table>
<thead>
<tr>
<th>Model</th>
<th>Inference</th>
<th>gender</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td></td>
<td>2.067(0.607, 3.527)</td>
<td>-0.031(-0.081, 0.020)</td>
</tr>
<tr>
<td>Full</td>
<td></td>
<td>2.237(0.859, 3.616)</td>
<td>-0.029(-0.077, 0.020)</td>
</tr>
<tr>
<td>Selected</td>
<td>Naive</td>
<td>0.466(0.330,0.603)</td>
<td>-0.020(-0.024,-0.016)</td>
</tr>
<tr>
<td></td>
<td>Selective</td>
<td>0.466(0.115,0.600)</td>
<td>-0.020(-0.024,-0.016)</td>
</tr>
</tbody>
</table>

**Table**: Coefficients and confidence intervals of gender (is female) and age obtained.

- Naive model is $Y_i = \mathbf{X}_i^T \gamma + T_i \mathbf{X}_i^T \beta + \epsilon_i$.
- Full model is $Y_i - \hat{\mu}_y(\mathbf{X}_i) = (T_i - \hat{\mu}_t(\mathbf{X}_i)) \mathbf{X}_i^T \beta + \epsilon_i$.
- Selected model is $Y_i - \hat{\mu}_y(\mathbf{X}_i) = (T_i - \hat{\mu}_t(\mathbf{X}_i)) \mathbf{X}_{i,\hat{M}}^T \beta_{\hat{M}} + \epsilon_i$.
- Except for “Selective inference”, all coefficients and confidence intervals are computed using \texttt{lm} in R.
Future directions

- Selective inference in general semiparametric setup.
- Target parameters defined by population instead of sample (ATT vs. SATT).
References


Proof Sketch

- Suppose we use \( \hat{\mu}_y = \mu_y \), then the pivot is exact for the following modified parameter

\[
\tilde{\beta}_M = \arg\min_{\alpha, \beta_M} \frac{1}{n} \sum_{i=1}^{n} \left( T_i - \hat{\mu}_t(X_i) \right)^2 \left( \Delta(X_i) - \alpha - X_{i,M}^T \beta_M \right)^2.
\]

- Show \( \| \tilde{\beta}_{\hat{M}} - \beta^*_M \|_\infty = o_p(n^{-1/2}) \).

- Replace \( \tilde{\beta}_{\hat{M}} \) by \( \beta^*_M \) and \( \mu_y \) by \( \hat{\mu}_y \) in the pivot, show the difference is \( o_p(1) \).

- The actual proof is much more technical (mainly because estimation error complicates the selection event).
Assumptions in the paper

Assumption

(Fundamental assumptions in causal inference) For $i = 1, \ldots, n$,

1. **Consistency of the observed outcome:** $Y_i = Y_i(T_i)$;
2. **Unconfoundedness of the treatment assignment:** $T_i \perp \perp Y_i(t) | X_i$, $\forall t \in T$;
3. **Positivity (or Overlap) of the treatment assignment:** $T_i | X_i$ has a positive density with respect to a dominating measure on $\mathcal{T}$. In particular, we assume $\text{Var}(T_i | X_i)$ exists and is at least $1/C$ for some constant $C > 0$ and all $X_i \in \mathcal{X}$.
Assumptions in the paper II

Assumption

(Accuracy of treatment model) \( \| \hat{\mu}_t - \mu_t \|_{\infty} = o_p(n^{-1/4}). \)

Assumption

The support of \( \mathbf{X} \) is uniformly bounded, i.e. \( \mathcal{X} \subseteq [-C, C]^p \) for some constant \( C \). The conditional treatment effect \( \Delta(\mathbf{X}) \) is also bounded by \( C \).

Assumption

(Accuracy of outcome model) \( \| \hat{\mu}_y - \mu_y \|_{\infty} = o_p(1) \) and \( \| \hat{\mu}_t - \mu_t \|_{\infty} \cdot \| \hat{\mu}_y - \mu_y \|_{\infty} = o_p(n^{-1/2}). \)
Assumptions in the paper III

Assumption

(Size of the selected model) For some constant $m$, $\Pr(|\hat{M}| \leq m) \to 1$.

Assumption

(Gram matrix) For all $M$ such that $|M| \leq m$, $E[X_i,MX_i,1,M] \succeq (1/C) I_{|M|}$.

The last two assumptions ensure $\|\tilde{\beta}_{\hat{M}} - \beta^*_\hat{M}\|_{\infty} = o_p(n^{-1/2})$. 
Assumptions in the paper IV

Assumption

(Truncation threshold) The truncation thresholds $L$ and $U$ satisfy

$$P\left( \frac{U(Y - \hat{\mu}_y) - L(Y - \hat{\mu}_y)}{\sigma \|\tilde{\eta}_M\|} \geq 1/C \right) \to 1.$$

Assumption

(Lasso solution)

$$P\left( \left| \left( \hat{\beta}_{\{1,\ldots,p\}}(\lambda) \right)_k \right| \geq 1/(C\sqrt{n}), \ \forall k \in \hat{M} \right) \to 1.$$

These two assumptions ensure the pivot is smooth enough.