

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

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- 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

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- 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:

$$Y_i(t) = \eta(\mathbf{X}_i) + t \cdot \Delta(\mathbf{X}_i) + \epsilon_i(t), \quad i = 1, \dots, n.$$

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

Basic assumptions:

Assumption

- 1 *Consistency of the observed outcome:* $Y_i = Y_i(T_i)$;
- 2 *Unconfoundedness:* $T_i \perp\!\!\!\perp Y_i(t) | \mathbf{X}_i, \forall t \in \mathcal{T}$;
- 3 *Positivity/Overlap:* $\text{Var}(T_i | \mathbf{X}_i = \mathbf{x})$ exists and is bounded away from 0 for all \mathbf{x} .

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

Naive linear modeling I

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

A straw man

Instead of the nonparametric model,

$$Y_i(t) = \eta(\mathbf{X}_i) + t \cdot \Delta(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

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

$$Y_i(t) = \gamma^T \mathbf{X}_i + T_i \cdot (\beta^T \mathbf{X}_i) + \tilde{\epsilon}_i, \quad i = 1, \dots, n.$$

Dismiss all insignificant interaction terms, then refit the model.

Naive linear modeling II

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

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 I

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- Acknowledge that the model is selected using the data.
- Model selection procedure:

$$\{\mathbf{X}_i, T_i, Y_i\}_{i=1}^n \mapsto \hat{\mathcal{M}} \quad (\text{data} \mapsto \text{a subset of covariates}).$$

- The target parameter $\beta_{\hat{\mathcal{M}}}^*$ is defined by $\hat{\mathcal{M}}$: $\mathbf{x}_{\hat{\mathcal{M}}}^T \beta_{\hat{\mathcal{M}}}^*$ is the “best linear approximation” of $\Delta(\mathbf{x})$ [Berk, Brown, Buja, Zhang, and Zhao, 2013].

Background: valid inference after model selection II

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- Two types of confidence intervals:

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

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

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

$$P\left(\left(\beta_{\mathcal{M}}^*\right)_j \in [D_j^-, D_j^+] \mid \hat{\mathcal{M}} = \mathcal{M}\right) \geq 1 - q, \forall \mathcal{M}.$$

Guarantees the control of false coverage rate (FCR, the average proportion of non-covering intervals among the reported) [Benjamini and Yekutieli, 2005].

Background: selective inference in linear models I

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

Suppose we have noisy observations of Δ :

$$Y_i = \Delta(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

- Submodel parameter

$$\beta_{\mathcal{M}}^* = \arg \min_{\alpha, \beta_{\mathcal{M}}} \sum_{i=1}^n \left(\Delta(\mathbf{X}_i) - \alpha - \mathbf{X}_{i,\mathcal{M}}^T \beta_{\mathcal{M}} \right)^2.$$

- Linear selection rule

$$\{\hat{\mathcal{M}} = \mathcal{M}\} = \{\mathbf{A}_{\mathcal{M}}(\mathbf{X}) \cdot \mathbf{Y} \leq \mathbf{b}_{\mathcal{M}}(\mathbf{X})\}.$$

- Example: Nonzero elements in the Lasso solution.

Background: selective inference in linear models II

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

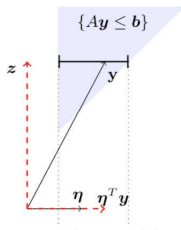
Future work

References

- Main result of Lee et al. [2016]:

$(\hat{\beta}_{\hat{\mathcal{M}}})_j | \mathbf{A}\mathbf{Y} \leq \mathbf{b}$ is truncated normal with mean $(\beta_{\hat{\mathcal{M}}}^*)_j$.

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



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

Eliminate the nuisance parameter

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- Back to the causal model (of the observables)

$$Y_i = \eta(\mathbf{X}_i) + T_i \cdot \Delta(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \dots, n.$$

- Problem: how to eliminate the nuisance parameter $\eta(\mathbf{x})$?

Robinson [1988]'s transformation

Let $\mu_y(\mathbf{x}) = E[Y_i | \mathbf{X}_i = \mathbf{x}]$ and $\mu_t(\mathbf{x}) = E[T_i | \mathbf{X}_i = \mathbf{x}]$, so $\mu_y(\mathbf{x}) = \eta(\mathbf{x}) + \mu_t(\mathbf{x})\Delta(\mathbf{x})$. An equivalent model is

$$Y_i - \mu_y(\mathbf{X}_i) = (T_i - \mu_t(\mathbf{X}_i)) \cdot \Delta(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \dots, n.$$

- The new nuisance parameters $\mu_y(\mathbf{x})$ and $\mu_t(\mathbf{x})$ can be directly estimated from the data.

Our complete proposal

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- Estimate $\mu_y(\mathbf{x})$ and $\mu_t(\mathbf{x})$ using machine learning algorithms (for example random forest).
- Select a model for effect modification by solving

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

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

$$\beta_{\hat{\mathcal{M}}}^* = \arg \min_{\alpha, \beta_{\hat{\mathcal{M}}}} \sum_{i=1}^n (T_i - \mu_t(\mathbf{X}_i))^2 (\Delta(\mathbf{X}_i) - \alpha - \mathbf{X}_{i, \hat{\mathcal{M}}}^T \beta_{\hat{\mathcal{M}}})^2.$$

Main result

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- Challenge: μ_y and μ_t are estimated (with error).

Assumption

Rate assumptions in Robinson [1988]:

$$\|\hat{\mu}_t - \mu_t\|_\infty = o_p(n^{-1/4}), \quad \|\hat{\mu}_y - \mu_y\|_\infty = o_p(1),$$
$$\|\hat{\mu}_t - \mu_t\|_\infty \cdot \|\hat{\mu}_y - \mu_y\|_\infty = o_p(n^{-1/2}).$$

Theorem

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

Simulation

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

Idealized estimation error

The true design and the true outcome were generated by

$$\mathbf{X}_i \in \mathbb{R}^{30} \stackrel{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{I}), Y_i \stackrel{i.i.d.}{\sim} N(\mathbf{X}_i^T \boldsymbol{\beta}, 1), i = 1, \dots, n,$$

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

Then the design and the outcome were perturbed by

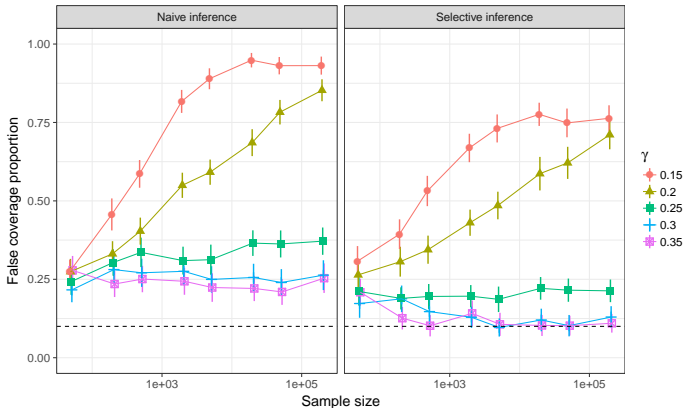
$$\mathbf{X}_i \mapsto \mathbf{X}_i \cdot (1 + n^{-\gamma} \mathbf{e}_{1i}), Y_i \mapsto Y_i + n^{-\gamma} \mathbf{e}_{2i},$$

where \mathbf{e}_{1i} and \mathbf{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

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- M. Visser, L. M. Bouter, G. M. McQuillan, M. H. Wener, and T. B. Harris. [Elevated C-reactive protein levels in overweight and obese adults.](#) *Journal of the American Medical Association*, 282(22): 2131–2135, 1999.
- 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

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

- T : obesity ($\text{BMI} \geq 25$).
- Y : C-reactive protein level.
- \mathbf{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.
- $n = 9677$, $p = 355$.
- $\mu_Y(\mathbf{x})$ and $\mu_T(\mathbf{x})$ are estimated by randomForest in R.
- By running our procedure, lasso found two effect modifiers: gender and age (no surprise!).

Real data example III

Effect modification

Qingyuan Zhao

Problem formulation

Selective inference: why and how

Selective inference for effect modification

Numerical examples

Future work

References

Model	Inference	gender	age
Naive		2.067(0.607, 3.527)	-0.031(-0.081, 0.020)
Full		2.237(0.859, 3.616)	-0.029(-0.077, 0.020)
Selected	Naive	0.466(0.330,0.603)	-0.020(-0.024,-0.016)
	Selective	0.466(0.115,0.600)	-0.020(-0.024,-0.016)

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

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

Future directions

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

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

References

- Y. Benjamini and D. Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.
- R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, et al. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.
- W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.
- J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.
- P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.
- X. Tian and J. Taylor. Asymptotics of selective inference. *Scandinavian Journal of Statistics*, to appear, 2017a.
- X. Tian and J. E. Taylor. Selective inference with a randomized response. *Annals of Statistics*, to appear, 2017b.
- M. J. van der Laan and S. Rose. *Targeted Learning*. Springer, 2011.
- M. Visser, L. M. Bouter, G. M. McQuillan, M. H. Wener, and T. B. Harris. Elevated C-reactive protein levels in overweight and obese adults. *Journal of the American Medical Association*, 282(22):2131–2135, 1999.

Proof Sketch

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

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

$$\tilde{\beta}_{\mathcal{M}} = \arg \min_{\alpha, \beta_{\mathcal{M}}} \frac{1}{n} \sum_{i=1}^n (T_i - \hat{\mu}_t(\mathbf{X}_i))^2 (\Delta(\mathbf{X}_i) - \alpha - \mathbf{X}_{i,\mathcal{M}}^T \beta_{\mathcal{M}})^2.$$

- Show $\|\tilde{\beta}_{\hat{\mathcal{M}}} - \beta_{\hat{\mathcal{M}}}^*\|_{\infty} = o_p(n^{-1/2})$.
- Replace $\tilde{\beta}_{\hat{\mathcal{M}}}$ by $\beta_{\hat{\mathcal{M}}}^*$ and μ_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 I

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

Assumption

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

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

Assumptions in the paper II

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

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

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

Assumption

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

Assumption

(Gram matrix) For all \mathcal{M} such that $|\mathcal{M}| \leq m$,
 $E[\mathbf{X}_{i,\mathcal{M}}\mathbf{X}_{i,\mathcal{M}}^T] \succeq (1/C)\mathbf{I}_{|\mathcal{M}|}$.

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

Assumptions in the paper IV

Effect
modification

Qingyuan
Zhao

Problem
formulation

Selective
inference: why
and how

Selective
inference for
effect
modification

Numerical
examples

Future work

References

Assumption

(Truncation threshold) The truncation thresholds L and U satisfy

$$P\left(\frac{U(\mathbf{Y} - \hat{\boldsymbol{\mu}}_y) - L(\mathbf{Y} - \hat{\boldsymbol{\mu}}_y)}{\sigma \|\tilde{\boldsymbol{\eta}}_{\mathcal{M}}\|} \geq 1/C\right) \rightarrow 1.$$

Assumption

(Lasso solution)

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

These two assumptions ensure the pivot is smooth enough.