FACULTY OF SCIENCES

MAY 16, 2024 AMERICAN CAUSAL INFERENCE CONFERENCE 2024

ASSUMPTION-LEAN QUANTILE REGRESSION

Supervisors: Prof. Stijn Vansteelandt and Prof. Christophe Ley

Georgi Baklicharov — georgi.baklicharov@ugent.be

GHENT UNIVERSITY

Modeling Interdisciplinarity Data science Applied Statistics

The modeling tradition

GHENT UNIVERSITY

THE MODELING TRADITION

Statistical modeling

- Models are central
- Recent critiques:
  (Breiman, 2001; Freedman, 2001; Robins & Rotnitzky, 2001; van der Laan, 2015; ...)
  - Occam's dilemma: simple & interpretable vs complex & plausible.
  - We need to make compromises $\Rightarrow$ misspecification and bias.
  - Model building $\Rightarrow$ bias and post-selection inference (Leeb & Pötscher, 2006; Dukes & Vansteelandt, 2020)

Algorithmic modeling

- Model misspecification is much less a concern.
- But focus is on prediction.
- Not aimed at explanation.
- No real uncertainty assessments.

The causal modeling culture increasingly builds on the algorithmic culture, instead targeting model-free estimands and providing valid uncertainty assessments.

How can we bridge these modeling cultures?

GHENT UNIVERSITY

Assumption-lean regression (1)

- That is what is achieved in a recent JRSS B discussion paper on assumption-lean modeling.

  Vansteelandt S, Dukes O. Assumption-lean inference for generalised linear model parameters (with discussion). JRSS-B 2022.

- Consider the semi-parametric structural quantile model

$$\underbrace{Q_\tau(Y^a|L)}_{Q_\tau(Y|A=a,L)} - \underbrace{Q_\tau(Y^0|L)}_{\text{unknown fct of } L} = \beta_\tau(L)a \quad \text{for all } a.$$

- Assume that adjustment for L suffices to control for confounding: $Y^a \perp\!\!\!\perp A|L$.
- Techniques for partially linear quantile models are relevant, but have limited utility:

  (Lee, 2003; Sun, 2005; Wu et al., 2010; Wu and Yu, 2014; Lv et al., 2015; Sherwood and Wang, 2016; Zhong and Wang, 2023)

  - computational demands;
  - challenges in high-dimensional applications (due to reliance on kernel weighting or splines);
  - biased inference when the model is wrong.

Assumption-lean regression (2)

- Because the model

$$Q_\tau(Y^a|L) - Q_\tau(Y^0|L) = \beta_\tau(L)a \quad \text{for all } a$$

  is deliberately kept simple, we will not assume it to hold, but use it to communicate our results.
- The real modeling is done through statistical / machine learning, results of which are projected and de-biased **in view of a specific estimand**.
- As such, we ensure that we are estimating a well-understood exposure effect and obtain valid inferences, even when the model is misspecified, and despite the use of machine learning.

Assumption-lean quantile regression

GHENT UNIVERSITY

Be clear about the estimand (1)

- A 'hygienic' analysis is clear about the estimand, even when models are used.
- For instance, with a binary randomized treatment A,
  we map $\beta_\tau(L)$ in model
  $$Q_\tau(Y^1|L) - Q_\tau(Y^0|L) = \beta_\tau(L)$$
  onto the model-free estimand
  $$\mathbb{E}\left\{Q_\tau(Y^1|L) - Q_\tau(Y^0|L)\right\},$$
  which is what we will estimate.
- This choice prevents that naïve interpretation as a 'difference between quantiles' would be misleading.
- In contrast, in standard (partially linear) quantile regression,
  it is unclear what we are estimating when the model is wrong.

Be clear about the estimand (2)

When A is not randomized, we may consider the same estimand, or generalize it to the weighted average:

$$\frac{\mathbb{E}[w(L)\left\{Q_\tau(Y^1|L) - Q_\tau(Y^0|L)\right\}]}{\mathbb{E}\left\{w(L)\right\}},$$

with

$$w(L) = \mathrm{P}(A = 1|L)\mathrm{P}(A = 0|L).$$

DEBIASED MACHINE LEARNING

GHENT
UNIVERSITY

A DEBIASED ESTIMATOR

- When $Y^a \perp\!\!\!\perp A | L$, the estimand can be identified as

$$\frac{\mathbb{E}\left(\{A - \mathbb{E}(A|L)\}\left[Q_\tau(Y|A,L) - \mathbb{E}\{Q_\tau(Y|A,L)|L\}\right]\right)}{\mathbb{E}\left[\{A - \mathbb{E}(A|L)\}^2\right]}$$

- Based on the estimand's efficient influence function,
  we construct the following debiased estimator

$$\frac{1}{n}\sum_{i=1}^{n} \frac{A_i - \hat{\mathbb{E}}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i - \hat{\mathbb{E}}(A_i|L_i)\}^2}\left[\hat{Q}_\tau(Y_i|A_i,L_i) - \hat{\mathbb{E}}\left\{\hat{Q}_\tau(Y_i|A_i,L_i)\middle|L_i\right\}\right] + \frac{1}{n}\sum_{i=1}^{n} \frac{A_i - \hat{\mathbb{E}}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}\{A_i - \hat{\mathbb{E}}(A_i|L_i)\}^2}\left[\frac{\tau - I\{Y_i \le \hat{Q}_\tau(Y_i|A_i,L_i)\}}{\hat{f}_{Y|A,L}(\hat{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i)}\right],$$

where the nuisance parameters are substituted by data-adaptive estimates (e.g., ML).

A Targeted Learning estimator

- Targeted learning 'simplifies' this by forcing the second line to give zero, which gives an asymptotically equivalent estimator.
- It does so by 'targeting' an initial estimator $\widetilde{Q}_\tau(Y|A,L)$ so that

$$\frac{1}{n}\sum_{i=1}^{n}\left\{A_i - \hat{\mathbb{E}}(A_i|L_i)\right\}\left[\frac{\tau - I\{Y_i \leq \widetilde{Q}_\tau(Y_i|A_i,L_i)\}}{\hat{f}_{Y|A,L}(\widetilde{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i)}\right] \approx 0.$$

- This is done by fitting the quantile regression model

$$\widetilde{Q}_\tau(Y_i|A_i,L_i) = \hat{Q}_\tau(Y_i|A_i,L_i) + \delta \cdot \frac{A_i - \hat{\mathbb{E}}(A_i|L_i)}{\hat{f}(\hat{Q}_\tau(Y_i|A_i,L_i)|A_i,L_i)}$$

- Next, we calculate the estimator as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i - \hat{\mathbb{E}}(A_i|L_i)}{\frac{1}{n}\sum_{i=1}^{n}(A_i - \hat{\mathbb{E}}(A_i|L_i))^2}\left[\widetilde{Q}_\tau(Y_i|A_i,L_i) - \hat{\mathbb{E}}\left(\widetilde{Q}_\tau(Y_i|A_i,L_i)\Big|L_i\right)\right].$$

Inference

- Inference is based on the efficient influence function after cross-fitting.
- We furthermore require the following terms to be $o_p(n^{-1/2})$:

$$\mathbb{E}\left[(\hat{Q}_\tau(Y|A,\boldsymbol{L}) - Q_\tau(Y|A,\boldsymbol{L}))^2\right],$$

$$\mathbb{E}\left[\left(1 - \frac{f(Q_\tau(Y|A,\boldsymbol{L})|A,\boldsymbol{L})}{\hat{f}(\hat{Q}_\tau(Y|A,\boldsymbol{L})|A,\boldsymbol{L})}\right)^2\right]^{1/2} \mathbb{E}\left[(\hat{Q}_\tau(Y|A,\boldsymbol{L}) - Q_\tau(Y|A,\boldsymbol{L}))^2\right]^{1/2},$$

$$\mathbb{E}\left[(\mathbb{E}(A|\boldsymbol{L}) - \hat{\mathbb{E}}(A|\boldsymbol{L}))^2\right]^{1/2} \mathbb{E}\left[\left(\mathbb{E}(Q_\tau(Y|A,\boldsymbol{L})|\boldsymbol{L}) - \hat{\mathbb{E}}(\hat{Q}_\tau(Y|A,\boldsymbol{L})|\boldsymbol{L})\right)^2\right]^{1/2},$$

$$\mathbb{E}\left[(\mathbb{E}(A|\boldsymbol{L}) - \hat{\mathbb{E}}(A|\boldsymbol{L}))^2\right] \qquad (\text{if } \beta_\tau \neq 0).$$

  - Weaker than standard parametric assumptions, but still non-negligible.
  - This is why our inferences are assumption-lean, rather than assumption-free

SIMULATION STUDIES

Simulation studies

- We considered inference for β_τ in

$$Q_\tau(Y^a|L) - Q_\tau(Y^0|L) = \beta_\tau a \quad \text{for all } a.$$

- L is 4-dimensional multivariate normal.
- 2 settings:
  - Binary exposure: $\mathbb{P}(A = 1|L) = \text{expit}(-0.5 + 0.2L_1 - 0.4L_2 - 0.4L_3 + 0.2L_4)$.
  - Continuous exposure: $A \sim \mathcal{N}(-0.5 + L_1 - 2L_2 - 2L_3 + L_4, 2^2)$.
- The outcome was generated according to

$$Y = 1 + A + \sin(L_1) + L_2^2 + L_3 + L_4 + L_3 \cdot L_4 + \epsilon,$$

  where $\epsilon \sim \text{Gamma}(k, \theta)$.
- Nuisance parameters are estimated using ‘grf’, ‘SuperLearner’ and ‘FKSUM’ R-packages.
- We contrast the proposal with an oracle quantile regression and a naive plug-in estimator.

Simulation Studies

Setting	estimator	$\tau = 0.5$				$\tau = 0.9$			
		bias	SD	SE	Cov	bias	SD	SE	Cov
	Oracle	-0.0017	0.19	0.20	96.6	-0.011	0.56	0.60	96.0
Bin.	Plugin	-0.70	0.12	0.015	0.1	-0.64	0.22	0.036	1.6
	TL-CF	0.012	0.22	0.25	97.2	0.14	0.68	0.63	91.4
	Oracle	-0.0013	0.035	0.036	95.6	0.0010	0.10	0.11	94.6
Cont.	Plugin	-0.17	0.064	0.016	0.5	-0.39	0.11	0.021	0.0
	TL-CF	-0.011	0.044	0.042	92.9	0.012	0.14	0.10	85.3

- Sample size $n = 500$, quantile τ, 1000 simulations
- Oracle: correctly specified QR
- Plugin: Naive plug-in estimator
- TL-CF: Targeted Learning with 5-fold cross-fitting

- bias: Monte Carlo bias
- SD: Monte Carlo standard deviation
- SE: averaged estimated standard error
- Cov: coverage of 95% CI

Conclusion

CONCLUSION

- Assumption-lean modeling aims to make statistical / causal analyses more hygienic, by being clear about what we are estimating when the model is wrong.
- Obtain valid inferences, despite the use of flexible data-adaptive / machine learning algorithms, even when the model is wrong.
- By focusing on conditional quantiles, we can
  - tackle continuous exposures,
  - make better patient-specific treatment decisions, and
  - study treatment effect heterogeneity.

References (1)

Hines, O., Dukes, O., Diaz-Ordaz K., and Vansteelandt, S. (2021). Demystifying statistical learning based on efficient influence functions. The American Statistician, 1-48.

van der Laan, M. J., & Rose, S. (2011). Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media.

Vansteelandt, S., & Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters (with discussion). Journal of the Royal Statistical Society - B, 84, 657-685.

Vansteelandt, S. (2021). Statistical modeling in the age of data science. Observational Studies, 7, 217-228.

Vansteelandt, S., Van Lancker, K., Dukes, O. & Martinussen, T. Assumption-lean Cox regression. Journal of the American Statistical Association, in press.

References (2)

Baklicharov, G, Ley, C., Gorasso, V., Devleesschauwer, B., Vansteelandt, S. (2024). Assumption-Lean Quantile Regression. *arXiv preprint arXiv:2404.10495*.