FACULTY OF SCIENCES

CAUSALITY IN EXTREMES, GENEVA, 2024

ASSUMPTION-LEAN QUANTILE REGRESSION

Stijn Vansteelandt, Ghent University

joint with Georgi Baklicharov and Christophe Ley

GHENT UNIVERSITY

The modeling tradition

Ghent University

THE STATISTICAL MODELING TRADITION

- The introduction of generalized linear(mixed) models, quantile regression, ... marked an enormous revolution in statistical data analysis:
  - it provided flexibility to study a wide range of scientific questions in an accessible manner,
  - allowed more rigorous adjustments to be made,
  - and helped getting rid of certain poor practices (e.g., dichotomizing variables)
- Even so, the statistical modeling tradition has been severely critiqued.

(Breiman, 2001; Freedman, 2001; Robins and Rotnitzky, 2001; van der Laan, 2015; ...)

Critiques to the statistical modeling tradition

(Vansteelandt S. Statistical modeling in the Age of Data Science. Observational Studies. 2021;7(1):217-28.)

- Occam's dilemma leaves us torn between using simple and interpretable, versus complex and plausible models.

  (Breiman , 2001)

- Inferring the whole data-generating mechanism is an overly ambitious undertaking.

  (Breiman, 2001)

- Even if we concentrate on parts of it, misspecification of the remaining parts may induce large bias.

  (Robins, 2000)

- Such misspecification can be difficult to diagnose.

  (Rubin, 1999)

- Attempts towards model building themselves introduce bias and make honest uncertainty assessments difficult to obtain.

  (Leeb and Pötscher, 2006; Dukes and Vansteelandt, 2020)

What about other modeling cultures?

- Model misspecification is much less a concern in the algorithmic modeling culture.
  - But it focuses on prediction,
    but is not aimed at explanation, and provides no real uncertainty assessments.
- The causal modeling culture increasingly builds on this culture,
  instead targeting model-free estimands and providing valid uncertainty assessments.
  - But not rarely over-simplifying the scientific question,
    or returning to traditional use of (causal) models.

How can we bridge these modeling cultures?

GHENT UNIVERSITY

ASSUMPTION-LEAN REGRESSION (1)

- That is what we achieve 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.

- Assume that adjustment for L suffices to control for confounding: $Y^a \perp\!\!\!\perp A \mid L$.
- Consider the semi-parametric structural quantile model

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

- 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 model

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

  is deliberately kept simple, we will not assume it to hold.
- 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 β_τ in model
  $$Q_\tau(Y^1|L) - Q_\tau(Y^0|L) = \beta_\tau$$
  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

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

- This weighting gives the stability desired for widescale practical use.
- Because it changes the target population, we provide similarly weighted summary statistics.
- In contrast, standard quantile regression
  - also weighs the data, but we have a poor understanding how the weighting is done;
  - mixes the effects of A and L when the model is wrong.

Be clear about the estimand (3)

For arbitrary A, these estimands generalize
to a least squares projection of the quantile difference

$$Q_\tau(Y^a|L) - Q_\tau(Y^{a^*}|L) \quad \text{onto} \quad a - a^*$$

for exposure values a and a^* randomly and independently drawn with the same value L (averaged over L).

DEBIASED MACHINE LEARNING

GHENT UNIVERSITY

A DEBIASED ESTIMATOR

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

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

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

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

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

A TARGETED LEARNING ESTIMATOR (TMLE)

- 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 of β_τ as

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

Assessing standard errors

- Uncertainty in data-adaptive estimates is difficult to quantify.
- But proposed estimator is not sensitive to it
  when nuisance parameter estimators converge at faster than n to the quarter rates.
- The variance of the estimator can therefore be estimated
  as 1 over n times the sample variance of the influence functions
  as if the nuisance parameters were given.

A FEW CAVEATS

- When flexible machine learning methods are used,
  sample-splitting/cross-fitting should be used.

  (Zheng & van der Laan, 2011; Chernozhukov et al., 2018)
- This removes additional bias due to overfitting.
- In order for the learners to converge sufficiently fast (at faster than n to the quarter rates), we also require assumptions like smoothness/sparsity.
- These are weaker than standard parametric assumptions,
  but are still non-negligible.
- This is why our inferences are assumption-lean,
  rather than assumption-free.

Simulation studies

GHENT UNIVERSITY

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**
Bin.	Oracle	-0.0017	0.19	0.20	96.6	-0.011	0.56	0.60	96.0
	Plugin	-0.70	0.12	0.015	0.1	-0.64	0.22	0.036	1.6
	TMLE-CF	0.012	0.22	0.25	97.2	0.14	0.68	0.63	91.4
Cont.	Oracle	-0.0013	0.035	0.036	95.6	0.0010	0.10	0.11	94.6
	Plugin	-0.17	0.064	0.016	0.5	-0.39	0.11	0.021	0.0
	TMLE-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
- TMLE-CF: TMLE 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

GHENT
UNIVERSITY

More hygienic (causal) analyses (1)

- The starting point of the 'causal roadmap'
  is the postulation of a causal estimand linked to the scientific question.
- This gets forgotten
  - when causal models are used (e.g., MSMs, SNMMs, target trials, ...);
  - when the use of overly simplistic estimands
    drifts researchers away from the scientific question (e.g., dichotomizing exposures).
- Assumption-lean modeling aims to make statistical / causal analyses more hygienic,
  by being clear about what we are estimating when the models is wrong.
- It does this by transporting the concept of a causal estimand to the broader modeling context.

More hygienic (causal) analyses (2)

- This focus on estimands may be viewed as undesirable.
  - It is needed to be open about the statistical analysis,
    just like openness about causal assumptions is central to causal inference.
- Also the focus on generic estimands may be view as less desirable.
  - It is needed to give statistical / causal analyses flexibility and accessibility to non-experts.
  - It prevents being overly ambitious in descriptive etiologic studies

    (where it is too ambitious to think about hypothetical interventions)

    and does not prevent more refined analyses.

Features

- The flexibility of standard regression

  (e.g., it readily handles continuous exposures).
- It overcomes Occam's dilemma by separating modeling to summarise from (data-adaptive) modeling to handle the curse of dimensionality.

  (Breiman, 2001)
- It prevents model misspecification bias by incorporating flexible modeling, machine learning.
- It avoids to extract information from modeling assumptions by working under the nonparametric model.
- It enables valid (post-selection) inference after using machine learning, variable selection, model selection.
- It enables (near) pre-specification of the entire analysis.
- It tries to avoid making strong extrapolations.
- It is 'simple' to obtain.

REFERENCES

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. (2022). Assumption-lean Cox regression. Journal of the American Statistical Association, 1-10.