Quantile Regression for Hypothesis Testing and Hypothesis Screening at the Dawn of Big Data

David H. Rehkopf

One implication of the application of generalized linear models in epidemiology is that only mean differences matter. Their exclusive use implies that investigators believe other parts of the distribution do not contribute information to our understanding of the relation between exposure and outcome (other than for an assessment of variance). In this issue of EPIDEMIOLOGY, Liu and colleagues[1] examine the distributional differences in the association of education with a cardiovascular risk score and with body mass index (BMI). The novel contribution of this work is the use of quantile regression to examine how the associations of education with cardiovascular risk score and BMI vary at different points in the distribution of these outcomes. The authors add to a small but growing literature in epidemiology using this approach.[2,3]

THE MEAN IS NOT THE (ONLY) MESSAGE

Liu and colleagues[1] provide evidence that we miss important aspects of the effects of an exposure if we rely exclusively on mean-based analysis. They find larger associations between education and cardiovascular risk at higher risk levels, primarily among women. These results suggest an alternative take on the thesis of Rose, who contrasts a "population-based" with a "high-risk" approach.[4] Liu and colleagues find instead that a population-level intervention can both shift the mean and have a greater impact on those at highest risk. This is an important contribution and supports other recent work taking a more empirical approach to assessing the benefits of targeted versus population-wide interventions.[5–7] This paper also provides an example of how a public health intervention strategy can decrease socioeconomic disparities while increasing overall population health.

The use of quantile regression is particularly critical when assessing associations for risk factors, biomarkers, or surrogate end points that do not have a linear relation with disease or mortality. Current evidence suggests a nadir in the BMI–mortality relation near a BMI of 25 kg/m^2, with increasing mortality at both lower and higher levels of BMI.[8] Assuming a causal relation, it would be best to have an intervention that increases BMI when it is below 25 and decreases BMI when it is above 25. Although the usual solution is to use a categorical outcome (eg, obesity, BMI $\geq$30), quantile regression allows a more detailed view that does not implicitly assume a step function of risk. Nonlinearities between risk factor or surrogate end point outcomes with disease suggest there could be better understanding of overall health benefits of interventions with closer attention to where in the distribution effects are greatest.

POTENTIAL LIMITATIONS OF THE APPLICATION OF QUANTILE REGRESSION

Given these advantages, and the fact that many have cautioned against overreliance on mean comparisons,[9] why hasn't the quantile approach been used more often? Although

From the Division of General Medical Disciplines, Stanford University School of Medicine, Stanford, CA.
Correspondence: David H. Rehkopf, 1265 Welch Road, Stanford, CA 94305. E-mail: drehkopf@stanford.edu.

ISSN: 1044-3983/12/2305-0665
DOI: 10.1097/EDE.0b013e318261f7be

lack of availability in common software packages may have hindered use in the past, issues with the speed of model fit and appropriate calculation of standard errors are no longer detriments.[10]

Should consumers of quantile regression results be concerned about bias from selective publication of results, given the inherently greater number of associations examined? The history of limited use in epidemiology and more common use in econometrics suggests this is not likely to be an issue. Investigators have rarely focused on particular quantiles of effect that best support their hypothesis, and instead typically present the spectrum of results at various percentiles of the dependent variable. Quantile regression has tended to facilitate an honest presentation of a fuller picture of evidence.

A further putative limitation is a concern that it just does not matter—associations infrequently differ across the distribution of the dependent variable. It is not possible to assess this from the literature, given that investigators may have fit quantile regression models and not report results when distributional impacts were similar. However, where results are reported, differential effects are frequent. The extent to which common substantial and meaningful differences in association might occur across the distribution will be apparent only with more widespread use of the method.

SINGLE VERSUS MULTIPLE ASSOCIATIONS

The use of quantile regression is an important step forward for epidemiology in terms of a fuller understanding of associations of variables, with relatively few drawbacks. Liu et al have appropriately used this as hypothesis testing based on a specific question. More generally, however, epidemiologists may consider this detailed description of data to be a critical part of hypothesis screening,[11] and further consider how these effects may differ in subgroups of predictor variables. The results of Liu et al[1] suggest that for many of the distributional effects, results are stronger for women than men. There are clear priors to examining sex difference in heart disease risk, and the findings are replicated across two data sets; thus, there is reason to believe they are not spurious. However, other groups may also be differently affected by education, including racial/ethnic groups.[12] In the case of the analysis by Liu et al, quantile regression stratified by sex in two data sets with two outcomes produces 96 parameters of association (Tables 2, 3). Examining the differential associations by four racial/ethnic groups as well, the number would jump to 384. Even assuming stability of estimates with a large sample size, interpretation becomes difficult. Nevertheless, we have strong priors that both of these factors are likely important in modifying the association of interest.

To obtain more detailed and useful descriptions of the data, it may be advantageous to use machine-learning tools, such as recursive partitioning, random forest and support vector machines that search for a best model fit including higher-order interactions and nonlinear relations between variables.[13,14] Such methods now underlie an approach to developing data-driven models of treatment in causal modeling,[15] but the details of the treatment models are typically not interpreted. Recent developments in machine learning have also produced hybrid approaches with generalized linear models to allow conditioning on known confounders followed by a search across a large number of covariates to build a model that identifies heterogeneity in effects.[16] Applications in epidemiology using machine learning include analysis of biomarker predictors of mortality,[17] Alzheimer disease,[18] and social and behavioral predictors of child obesity.[19] As the epidemiologic literature over time becomes richer—providing investigators with more priors on potential predictors and moderators of importance—the number of effect estimates produced by stratification can grow beyond interpretation and, unfortunately, create ample opportunities for selective emphasis of findings. On the other hand, machine learning with internal cross-validation can be used to determine the most robust and substantial relations. A further advantage of machine learning is a transparent and replicable approach for examining nonlinearity and higher-order interactions between covariates that further diminishes the possibilities of selective emphasis of results.

With increasingly detailed and large data sets (the dawn of "Big Data"),[20] and more opportunities for out-of-sample cross validation, epidemiologists are poised to be able to examine more detailed relations among variables with quantile regression and machine learning methods. Although these more detailed descriptions of data have seen little use, thus far, in epidemiology, it is quite possible that these approaches may reinvigorate the methodologically stagnant endeavors of descriptive epidemiology, allowing for better informed priors for testing hypotheses using causal models.

ABOUT THE AUTHOR

DAVID REHKOPF is an Assistant Professor of Medicine in the Division of General Medical Disciplines at Stanford University School of Medicine. His research interest is in understanding the role of socioeconomic factors in chronic disease and in the application of underutilized statistical methods that allow for interaction and nonlinear relationships.

ACKNOWLEDGMENTS

I thank Jennifer Ahern and Mark Cullen for helpful comments on an earlier version of this commentary.

REFERENCES

1. Liu SY, Glymour M, Kawachi I. Education and inequalities in risk scores for coronary heart disease and body mass index: evidence for a population strategy. *Epidemiology*. 2012;23:657–664.
2. Burgette LF, Reiter JP, Miranda ML. Exploratory quantile regression

with many covariates: an application to adverse birth outcomes. *Epidemiology*. 2011;22:859–866.
3. Chen Q, Garabrant DH, Hedgeman E, et al. Estimation of background serum 2,3,7,8-TCDD concentrations by using quantile regression in the UMDES and NHANES populations. *Epidemiology*. 2010;21(suppl 4): S51–S57.
4. Rose G. Sick individuals and sick populations. *Int J Epidemiol*. 1985; 14:32–38.
5. Zulman DM, Vijan S, Omenn GS, Hayward RA. The relative merits of population-based and targeted prevention strategies. *Milbank Q*. 2008; 86:557–580.
6. Frohlich KL, Potvin L. Transcending the known in public health practice: the inequality paradox: the population approach and vulnerable populations. *Am J Public Health*. 2008;98:216–221.
7. Ahern J, Jones MR, Bakshis E, Galea S. Revisiting Rose: comparing the benefits and costs of population-wide and targeted interventions. *Milbank Q*. 2008;86:581–600.
8. Durazo-Arvizu R, McGee D, Li Z, Cooper R. Establishing the nadir of the body mass index-mortality relationship: a case study. *J Am Stat Assoc*. 1997;92:1,312–9.
9. Koenker R. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *J Econom*. 2000;95:347–374.
10. Koenker R, Hallock KF. Quantile regression. *J Econ Perspect*. 2001; 15:143–156.
11. Rothman K, Greenland S, Lash T. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
12. Kimbro RT, Bzostek S, Goldman N, Rodriguez G. Race, ethnicity, and the education gradient in health. *Health Affairs*. 2008;27:361–372.
13. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
14. Breiman L. Statistical modeling: the two cultures. *Stat Sci*. 2001;16: 199–231.
15. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article25.
16. Hothorn T, Hornik K, Zeileis A. Party: A Laboratory for Recursive Part(y)itioning. R package version 0.9–0 ed2006.
17. Gruenewald TL, Seeman TE, Ryff CD, Karlamangla AS, Singer BH. Combinations of biomarkers predictive of later life mortality. *Proc Natl Acad Sci U S A*. 2006;103:14158–14163.
18. Hu WT, Chen-Plotkin A, Arnold SE, et al. Novel CSF biomarkers for Alzheimer's disease and mild cognitive impairment. *Acta Neuropatol.* 2010;119:669–678.
19. Rehkopf DH, Laraia BA, Segal M, Braithwaite D, Epel E. The relative importance of predictors of body mass index change, overweight and obesity in adolescent girls. *Int J Pediatr Obes*. 2011;6: e233–242.
20. Lynch C. Big data: how do your data grow? *Nature*. 2008;455:28–29.