Abstract Title Page

Title:

A Bayesian Semiparametric Multivariate Causal Model, With Automatic Covariate Selection and For Possibly-Nonignorable Missing Data

Authors:

Karabatsos[1], G., and Walker[2], S.G.

[1]University of Illinois-Chicago, College of Education
[2]Kent University, Canterbury, United Kingdom, Institute of Mathematics, Statistics and Actuarial Science.

Abstract Body

Background/context

Causal inference is central to educational research, where in data analysis the aim is to learn the causal effects of educational treatments on academic achievement, to evaluate educational policies and practice. Compared to a correlational analysis, a causal analysis enables policymakers to make more meaningful statements about the efficacy of educational treatments.

A causal effect is a comparison of the potential outcome (e.g., literacy achievement) of a subject in response to receiving a control treatment (e.g., old teaching method), against the potential outcome of the same subject in response to receiving an active treatment (e.g., a new teaching method) (Neyman, 1923; Rubin, 1974, 1977). The fundamental problem of causal inference is that, at a given time, each subject can be exposed to only one of the treatments (Holland, 1986). Therefore, causal inference can be approached as a problem in multivariate regression with missing potential outcome (dependent-variable) data, given a set of covariates, where only one of the potential outcomes is observable from each subject, and where a primary aim is to impute plausible values of the missing potential outcomes to infer the causal effects for each subject (Rubin, 1978). Moreover, when treatment assignment probabilities are unknown for a set of subjects, as in an observational (non-randomized) study, the multivariate regression model can be expanded to jointly include a multinomial regression model that describes the distribution of the treatment assignments conditional on covariates, and conditional on potential outcomes to account for confounded treatment assignments (Rubin, 1978).

Causal inference becomes inaccurate whenever data violate certain assumptions that are often made in practice, including: (1) the usual assumption of no outliers in the potential outcomes, (2) the typical assumptions that the treatment assignments have no outliers, no hidden bias (e.g., Rosenbaum, 2002), no confounding, and satisfy the Stable Unit Treatment Value Assumption (SUTVA; Cox, 1958); (3) the usual assumption that the missing data values are either missing-at-random (MAR) or missing-completely-at-random (MCAR) (Little & Rubin, 2002; Ibrahim, Chen, Lipsitz, & Herring, 2005), and (4) the usual assumption that parameter estimation requires no penalty for the absolute size of regression coefficients. However, it is reasonable to believe that at least one of these assumptions is invalid for many data sets of educational research, where it is common to find outliers in the potential outcomes and in the treatment assignments, it is common to find outliers, hidden bias, confounding, and interference violations of SUTVA because students within a classroom or school interact (Rubin, 1990), and where it is common to find that missing data are non-ignorable (non-random) instead of MAR or MCAR. Finally, while many educational data sets contain data on a large number of covariates, such a large number can lead to high covariances in the unpenalized regression coefficient estimates, causing poor predictions (e.g., Hastie, et al. 2001, Ch. 3.4). Moreover, stepwise approaches to variable (covariate) selection, often used in the practice of regression, can be problematic (Pohlmann 1979; Adams, 1991; Roecker, 1991; Freedman, et al., 1992; Derksen & Keselman, 1992).

Purpose / objective / research question / focus of study:

To address the four open issues of causal modeling, we introduce a Bayesian semiparametric causal model, which provides a semiparametric approach to the full Rubin (1978) Causal Model.

The causal model includes, as appropriate, a multivariate-normal regression model for continuous-valued potential outcomes, or a multivariate-probit model for discrete-valued potential outcomes, conditional on a set of possibly-many covariates. Interference violations of SUTVA can be addressed through the specification of additional potential outcomes that reflect both the treatment received by a subject and the treatments received by other subjects. When the treatment assignments are non-ignorable, as in a non-randomized, observational study, the causal model jointly includes a multinomial probit model to describe the distribution of binary or multi-valued (e.g., dosage) treatment assignments (as appropriate), which is also conditioned on a set of possibly-many covariates, including potential outcomes as covariates to account for any confounding in the treatment assignments. Moreover, for these two joint multivariate regression models, our causal model specifies a stick-breaking prior distribution (Ishwaran & James, 2001) for the mixing distribution of subject-level random intercepts and variances, to provide a flexible nonparametric mixture of multivariate normal regression models for the joint distribution of potential outcomes and treatment assignments. This, in turn, provides robust causal inferences by capturing any multi-modalities, outliers, skewness, heavy-tail behavior, hidden bias, and extra correlation in this joint distribution. The stick-breaking prior distribution is a general type prior that includes other important nonparametric priors as special cases, including the Dirichlet Process prior (Ferguson, 1973) and the two-parameter Pitman-Yor (1997) process.

To provide a computationally-efficient basis for multiply-imputing plausible values for the missing potential outcomes, covariates, and treatment assignment data that are either randomly-missing (MCAR or MAR) or nonignorably missing, the semiparametric causal model specifies a multivariate normal regression model for the covariate distribution, and specifies multivariate probit binary regression model for the recording mechanism that describes the joint distribution of missing-value indicators for all variables having any missing data values, given possibly many covariates. Each of these multivariate models is defined by a sequence of univariate regressions (Ibrahim, Lipsitz, & Chen, 1999). Multiple-imputation of plausible values for the missing data points is achieved by repeatedly sampling from the posterior predictive distribution of the full semiparametric causal model.

Thus, the full Bayesian semiparametric causal model can describe the joint distribution of potential outcomes, treatment assignments, covariates, and missing-value indicators, using four multivariate regression models that are each conditioned upon possibly-many covariates. To address the fourth issue that is commonly posed by a large number of covariates, the causal model specifies multiple-shrinkage prior to perform penalized estimation of all the slope coefficients in the model. Specifically, this prior assigns to each slope coefficient a zero-mean normal prior with variance assigned a multinomial hyper-prior supporting values ranging from near zero to a very large number. As a consequence, whenever a covariate is an irrelevant predictor of a dependent variable for a set of data, the posterior distribution of the variance concentrates near zero, causing the covariate's slope coefficient to shrink towards zero. The multiple-shrinkage prior provides a coherent, model-based approach to variable selection that automatically identifies important predictors of dependent variables in the posterior distribution, while ensuring stable prediction. To complete the specification of the full Bayesian semiparametric causal model, in addition to the specification of the multiple-shrinkage priors on all slope coefficients, an inverse-gamma prior is specified for each of the error-variance parameters of the sub-model for the covariate distribution. Also, a prior can be specified for the

baseline hyper-parameters of the stick-breaking prior distribution. In practice, given a set of data, inference of the full posterior distribution of the model is possible through the use of existing Gibbs sampling methods for linear mixed models, along with modern Gibbs-sampling methods for stick-breaking models (Walker, 2007). Metropolis-Hastings algorithms can be used to sample the posterior predictive distribution of the causal model, to multiply-impute values of the missing data, which among other things would enable the inference of causal effects.

The paper presents our semiparametric causal model in full detail. We then illustrate this model through the analysis of data from the Progress In International Reading Literacy Study (PIRLS), to infer the causal effects of a writing instructional treatment on the reading performance of low-income students. This analysis is performed in a typical context of an observational study where SUTVA is potentially violated by the interference of subjects within each classroom, with many covariates describing the student, teacher, classroom, and school, where hidden bias and confounding can be present, and where there are missing covariate, treatment assignment, and potential outcome data, that can either be randomly (MCAR or MAR) or nonignorably missing.

Setting:

The setting of the observational study deals with 28 4th-grade classrooms from a national sample of 21 low-income U.S. schools, where all students received either reduced or free lunch during year 2006. These schools had an average enrollment of 554.6 (S.D.=237.4), ranging from small to large enrollments (min=153, 25%ile=400, 50%ile=495, 75%ile=622, max=1030).

Population / Participants / Subjects:

The subjects of the study are a sample of 565 economically-disadvantaged 4th grade students from the 28 classrooms, 49.6% of whom are female, 86% were 9 to 10 years old and 14% were 11 to 13 years old. Each of the 28 classrooms had between 17 and 31 students. Twelve of these classrooms had between 5% and 33% English-language learners. Almost all classrooms had between 5% to 100% students needing remedial instruction, and about one-fourth of classrooms had at least 50% remedial students. Also, 86% of the students had teachers with between 1 to 4 years of 4th grade teaching experience (38.6 % had 1 year of experience) while the remaining teachers had between 6 to 11 years of experience, 56.6% of students had teachers with a bachelor's degree, and the remaining teachers had a higher degree.

Intervention / Program / Practice:

The active treatment is defined as the student instructed to write something (e.g., essay) after reading in at least almost every lesson, while the control treatment is defined by the student being instructed to write less frequently. In total, 36.5% of all students received the active treatment, 61.6% of the students received the control treatment, and the values of the treatment assignments were missing for 1.9% of the students. Also, in 6 of the 28 classrooms, more than 50% of the students received the active treatment.

Research Design:

The research design involves the analysis of secondary data made available most recently for

year 2006 by PIRLS. These data provide an observational study of the causal effects of the treatments, because the treatments were not randomly assigned to the students and the treatment assignment probabilities are unknown. These probabilities can be estimated from the PIRLS data through the specification of a multinomial model for the treatment assignment mechanism, in the full semiparametric causal model.

Data Collection and Analysis:

The secondary PIRLS data were obtained from http://timss.bc.edu/pirls2006/user_guide.html. The potential outcome variable was a standardized score on a literacy exam. For each student, the treatment variable was coded 2 if a student and more than half-of the classmates were frequently instructed to write something after reading, coded 1 if a student and less than half of the classmates were frequently instructed to write something after reading, and coded 0 if the student was not frequently instructed to write after reading. This coding scheme accounts for potential violations of SUTVA that may arise from interference of students within a classroom (Rubin, 1990), and there are three potential outcomes defined for each student (notated by $Y(2)$, $Y(1)$, and $Y(0)$), with only one observable potential outcome. The potential outcomes are modeled by a tri-variate normal regression model with three random intercepts and three random variances per student, while the three-category treatment assignments are treated as a dependent variable in a multinomial probit regression model with two random intercepts per student. For each of these two regression models, there were 80 covariates describing the student, classroom, teacher, and school, each of which were standardized to have mean zero and variance 1, to facilitate interpretation. The student-level covariates included gender, age, index of reading attitudes, index of reading self-concept, and index of student safety. The class-level covariates include 28 indicators of the classroom, class size, the percentage of students in classroom understanding English, the percentage of English language learners in classroom, and the percentage of students in classroom needing remedial reading instruction. The teacher-level covariates include the number of years taught 4th grade, an indicator of whether or not the teacher studied reading theory as part of her training, time spent in seminars for teaching reading, time spent on reading books relating to teaching reading, an indicator of whether the teacher studied remedial reading instruction as part of her training, the teachers level of formal education, an indicator of whether or not the teacher seeks help from parents of students who are behind in their reading lessons, the number of classroom hours per week spent on reading instruction, an index of teacher career satisfaction, and an index of how much the teacher assigns reading for homework. The school-level covariates include 21 indicators of the school, total student enrollment, and indices of student tardiness, student absenteeism, classroom disturbance, students' desire to do well, percentage of students with early literacy skills, an indices of principal's perception of school safety, school climate, availability of school resources, and home-school involvement. Additionally, the multivariate normal regression model included potential outcomes as covariates to account for any correlation among the three potential outcomes, and the joint, multinomial probit regression model included all potential outcomes as covariates to account for any confounding in the observed treatment assignments. Moreover, a relatively non-informative, stick-breaking prior distribution was specified for the mixing distribution of the student-level random intercepts and random variances.

While each of the 565 students had missing values for two out of the three potential outcomes, 1.9% had missing values for the student-level treatment assignments, between 2.5% and 6.7%

had missing values for 11 covariates, and in one covariate (number of classroom hours per week spent on reading instruction) 20.7% of the values were missing. To enable the multiple-imputation of all missing values that are either randomly-missing (MCAR or MAR) or nonignorably missing, the semiparametric causal model was expanded to jointly include a multivariate regression model for the covariate distribution, and to jointly include a multivariate probit binary probit regression model for the recording mechanism describing the joint distribution of the missing-value indicators, for all variables containing missing values. Finally, to enable automatic variable selection, a multiple-shrinkage prior was assigned to all slope coefficients of all the four multivariate regression models describing the joint distribution of potential outcomes, treatment assignments, covariates with missing values, and the missing-value indicators for variables with missing values. Also, each variance parameter of the covariate distribution was assigned a non-informative inverse-gamma prior.

Findings / Results:

All results are based on generating 20,000 samples from the posterior distribution of the causal model, obtained by running the sampling algorithm for 130,000 iterations, discarding the first 70,000 burn-in samples, and retaining every third of the remaining 60,000 samples. Figure 1 and Table 1 present the estimated posterior predictive densities of the two causal effects, for the 565 students. Among other things, these densities show that the causal effect $Y(1) - Y(0)$ is significantly positive, the causal effect $Y(2) - Y(0)$ is not significantly different from zero, both densities are rather heavy-tailed, and the density of $Y(1) - Y(0)$ bimodal. Also, the differences between the two causal effect densities seem to indicate that the causal model is accounting for interference violations of SUTVA. Figures 2 and 3 present the posterior estimates of the median causal effects of students grouped by different categories of student-level, classroom-level, teacher-level, and school-level variables. Tables 2 and 3 present the estimate of the posterior distribution of the random intercepts and variances, and the posterior estimate of their correlation matrix. The positive standard deviations of the random intercepts ($\beta_{00}, \beta_{01}, \beta_{02}$) and of the random variances ($\sigma_0^2, \sigma_1^2, \sigma_2^2$) indicate that the semiparametric causal model is accounting for extra sources of variation in the potential outcomes due to unrecorded covariates and due to heteroscedasticity, respectively. The positive standard deviations of the random intercepts of the multinomial regression model with 83 predictors for the treatment assignments, (λ_{01}, λ_{02}), indicate that the causal model is accounting for the presence of hidden bias. Also, this model showed that the potential outcomes were significant predictors of the treatment assignments, meaning that the causal model is accounting for confounding in the treatment assignments. Also, the 17-variate model for the recording mechanism, with each variate assigned between 85-97 predictors, showed that the missing-values of the covariate, potential outcome, and treatment assignments significantly predicted their corresponding missing-data indicator variable. This means that the causal model is accounting for non-ignorable missing data.

Conclusions:

We introduced, recommend, and illustrated a Bayesian semiparametric causal model, and showed that it is a practical model that circumvents typical assumptions in causal modeling that can be violated in a typical data set arising from educational research. We look forward to future applications of the model, to further understand the efficacy of various educational treatments.

Appendices

Appendix A. References

Adams, J. (1991). A computer experiment to evaluate regression strategies. In *Proceedings of the Computational Statistics Section, American Statistical Association* (p. 55-62). American Statistical Association.

Cox, D. (1958). *The planning of experiments*. New York: John Wiley.

Derksen, S., & Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45,* 265-282.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1,* 209-230.

Freedman, L., Pee, D., & Midthune, D. (1992). The problem of underestimating the residual error variance in forward stepwise regression. *The Statistician, 41,* 405-412.

Hastie, T., Tibshiriani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer-Verlag.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-960.

Ibrahim, J., Chen, M.-H., Lipsitz, S., & Herring, A. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association, 100,* 332-346.

Ibrahim, J., Lipsitz, S., & Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Series B, 61,* 173-190.

Ishwaran, H., & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association, 96,* 161-173.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data (second edition)*. New York: Wiley.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Annals of Agricultural Science; Translated in Statistical Science* 1990, *5*, 465-472.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability, 25,* 855-900.

Pohlmann, J. (1979). *Controlling the Type I error rate in stepwise regression analysis.* (Tech. Rep. No. ED171746). Education Resources Information Center (ERIC).

Roecker, E. (1991). Prediction error and its estimation for subset-selected models. *Technometrics, 33,* 459-468.

Rosenbaum, P. (2002). *Observational studies (2nd edition)*. New York: Springer-Verlag.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688-701.

Rubin, D. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1-26.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34-58.

Rubin, D. (1990). Neyman (1923) and causal inference in experiments and observational studies.

Statistical Science, 5, 472-480.
Walker, S. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation, 36*, 45-54.

Appendix B. Tables and Figures

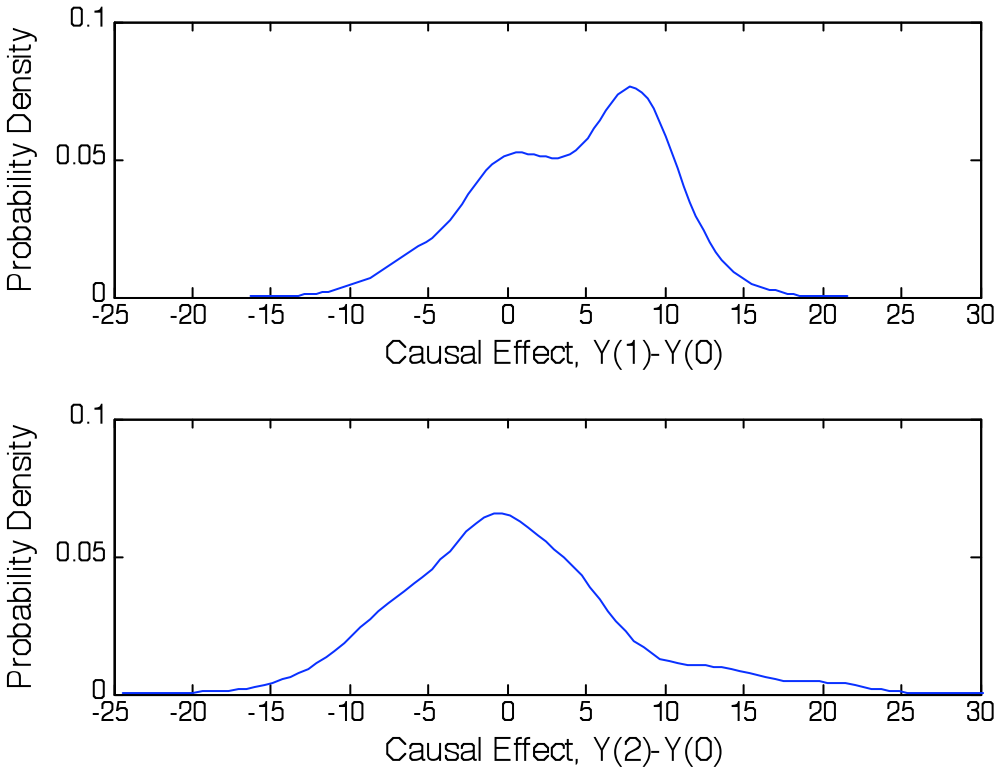

Figure 1: Posterior mean estimates of the marginal density of causal effects, in comparisons of potential outcomes Y(1) vs. Y(0), and Y(2) vs. Y(0).

Causal Effect:	$Y(1) - Y(0)$	$Y(2) - Y(0)$
Mean	4.13	.23
Median	4.19	–.08
95% Credible Interval of Median	(3.45,4.93)	(-3.88,2.25)
Number of 565 students significantly > 0	232	45
Number of 565 students significantly < 0	45	2
Mid-Quartile (MQ)	4.19	0.00
5%ile	–6.89	–14.30
25%ile	–.29	–5.71
75%ile	8.67	5.71
95%ile	14.92	16.34
S.D.	6.65	9.21
Inter-Quartile Range (IQR)	8.95	11.41
Skewness = (Median–MQ)/(2*IQR)	.00	–.00
Left tail size = (5%ile–MQ)/(2*IQR)	Medium, –.62	Medium, –.62
Right tail size = (95%ile–MQ)/(2*IQR)	Medium, .60	Medium, .75

Table 1: Posterior mean estimates of the causal effect distributions, and the 95% posterior credible interval of the median causal effect. A number of students having significant causal effects is based on a 95% posterior credible interval of the student causal effect.

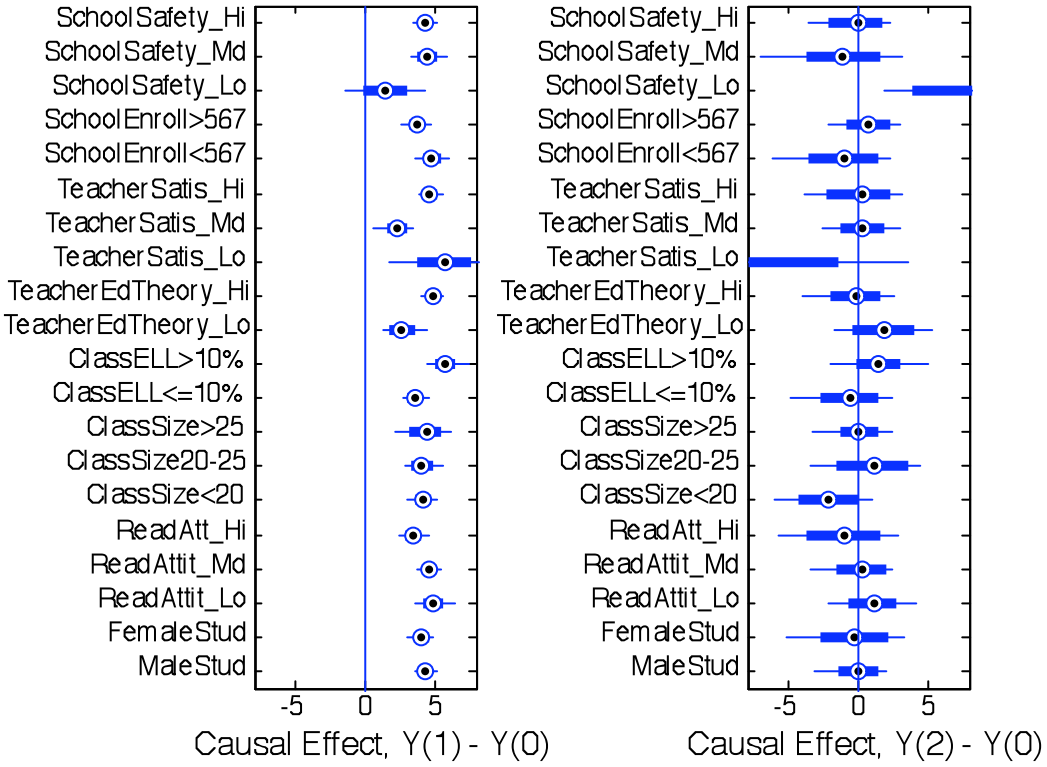

Figure 2: Posterior mean, 25%ile, 75%ile, and 95% credible interval estimates of median causal effects, of students grouped by different categories of student-level, classroom-level, teacher-level, and school-level variables. A vertical line that overlaps with zero indicates an insignificant causal effect for a group of students.

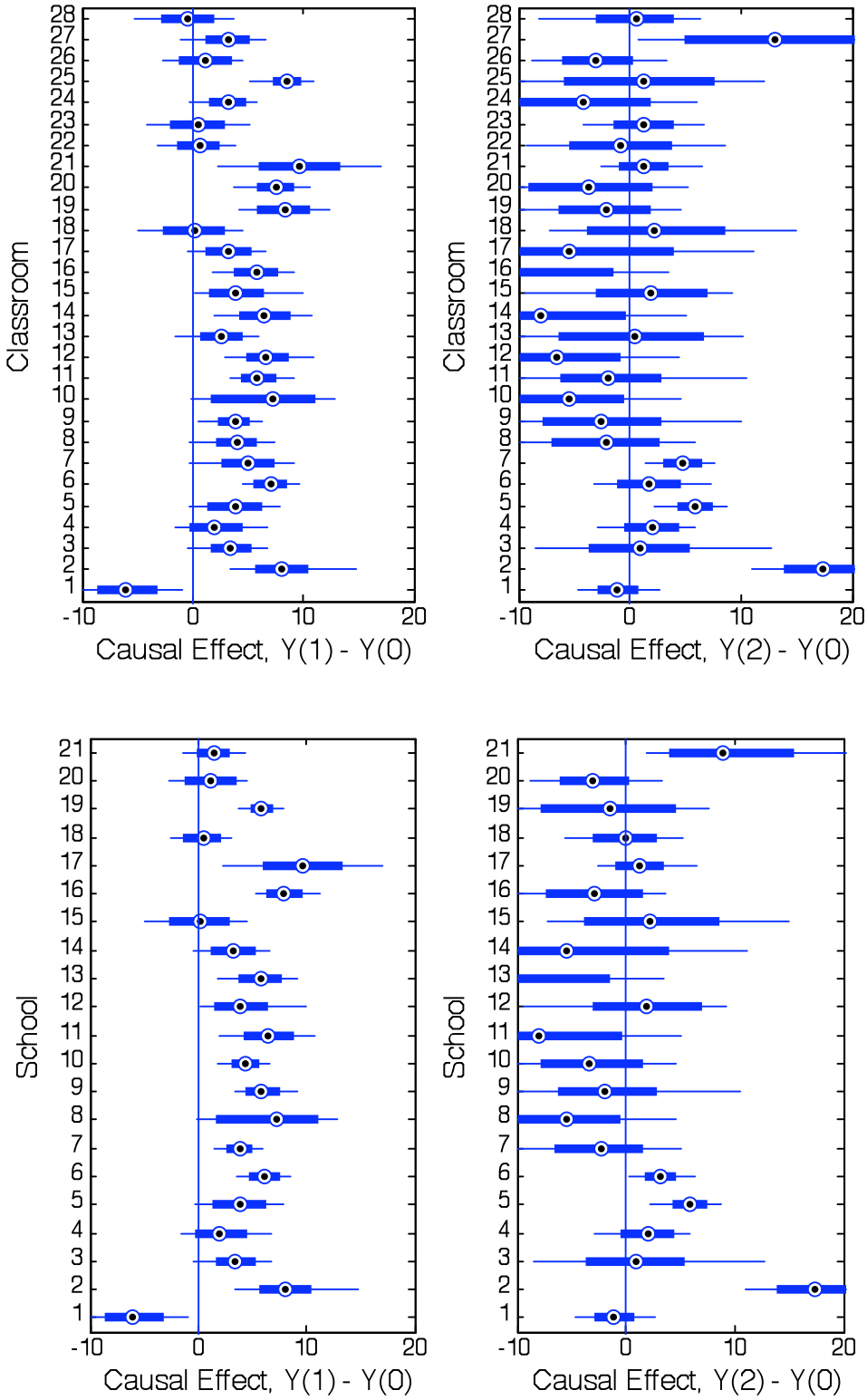

Figure 3: Posterior mean, 25%ile, 75%ile, and 95% credible interval estimates of median causal effects, for students grouped by classroom (top plots) and grouped by school (bottom plots). A vertical line overlapping with zero indicates an insignificant causal effect for a group of students.

$$\beta_{00} \quad \beta_{01} \quad \beta_{02} \quad \sigma_0^2 \quad \sigma_1^2 \quad \sigma_2^2 \quad \lambda_{01} \quad \lambda_{02}$$

Mean	41.72	6.32	7.30	64.06	27.01	1.78	12.97	–20.49
Median	45.31	6.10	7.78	68.91	27.72	1.63	14.70	–22.29
Mid-Quartile	45.34	6.51	7.70	68.83	27.39	1.58	14.21	–22.08
5%ile	2.79	1.71	2.23	1.21	1.23	.30	2.03	–24.50
25%ile	44.92	4.42	6.34	65.23	18.71	.87	12.38	–23.22
75%ile	45.76	8.60	9.05	72.43	36.08	2.28	16.04	–20.94
95%ile	46.40	12.68	11.77	77.68	45.02	4.40	17.14	–2.77
S.D.	12.75	4.35	4.06	19.30	11.88	1.20	5.63	6.93
IQR	.85	4.18	2.71	7.20	17.37	1.41	3.66	2.28
Skewness	–.01	–.05	.01	.01	.01	.02	0.07	–.05
Left tail size	–25.06 (long)	–.57 (med)	–1.01 (long)	–4.70 (long)	–.75 (med)	–.45 (med)	-1.66 (long)	–.53 (med)
Right tail size	.62 (med)	.74 (med)	.75 (med)	.62 (med)	.51 (med)	1.00 (long)	.40 (short)	4.24 (long)

Table 2: Posterior estimates of the marginal distributions of random intercepts ($\beta_{00}, \beta_{01}, \beta_{02}$) and random variances ($\sigma_0^2, \sigma_1^2, \sigma_2^2$) of the potential outcomes, and of the random intercepts (λ_{01}, λ_{02}) of the treatment assignments.

	β_{00}	β_{01}	β_{02}	σ_0^2	σ_1^2	σ_2^2	λ_{01}
β_{00}							
β_{01}	.42						
β_{02}	.55	.09					
σ_0^2	.95	.42	.57				
σ_1^2	.65	.66	.20	.63			
σ_2^2	.11	–.30	.30	.13	–.37		
λ_{01}	.71	.45	.27	.68	.73	–.17	
λ_{02}	–.86	–.38	–.51	–.85	–.62	–.10	–.67

Table 3: Posterior correlation matrix of the random intercepts and random variances.