Abstract Title Page

Title: Bayesian Unimodal Density Regression for Causal Inference[1]

Author(s): George Karabatsos, Associate Professor of Educational Psychology, University of Illinois-Chicago

Stephen G. Walker, Professor of Statistics, University of Kent, Canterbury

May 1, 2011

[1] This research is supported by the Chicago Teacher Partnership Project (CTPP) grant U336S090013, from the U.S. Department of Education Teacher Quality Partnership program.

Abstract Body

Background:

Karabatsos and Walker (2011) introduced a new Bayesian nonparametric (BNP) regression model. Through analyses of real and simulated data, they showed that the BNP regression model outperforms other parametric and nonparametric regression models of common use, in terms of predictive accuracy of the outcome (dependent) variable. The other, outperformed, regression models include random-effects/hierarchical linear and generalized linear models, when the random effects were assumed to be normally-distributed (Laird & Ware, 1982; Breslow & Clayton 1993), and when the random effects were more generally modeled by a nonparametric, Dirichlet process (DP) mixture prior (Kleinman & Ibrahim, 1998a,1998b).

Meanwhile, typical applications of causal inference focus on how a treatment causally affects the mean of the outcome, through the use of regression model that assume symmetrically-distributed errors. However, in many applications, it may also be of interest to investigate how the treatment causally changes other aspects of the outcome distribution, such as the median (for robustness), the 10th percentile to study treatment effect on lower-achieving students, or even the entire outcome distribution (density). Also, the symmetric distribution assumption is almost always violated by real data, and such a violation can decrease the accuracy of causal inferences.

Purpose of Study:

We argue that the new BNP regression model provides a novel, richer, and more valid approach to causal inference, which allows the researcher to investigate how treatments causally change the entire distribution (density) of (potential) outcomes, including not only the mean, but also other features of the outcome variable, such as quantiles (e.g., median, 10th percentile), and the variance. We illustrate the BNP model through the analysis of observational data, to estimate the causal effect of exposure to excellent high school math education (versus non-exposure, the control), on ACT math achievement. In the data analysis, we also compare the predictive accuracy of the new BNP model against other regression models. These other models assume symmetric distributions for the outcomes, and for the inverse-link function of the propensity score model (when specified), and have been recommended for causal inference from observational data.

The other models include the normal linear regression model, having one interaction between (1) subject (pre-treatment) covariates, (2) treatment indicators, and (3) indicators of ≥ 5 matched groups of subjects, formed either by subclassification (Rosenbaum & Rubin, 1984) or optimal full matching (Hansen & Klopfer, 2006; Rosenbaum,1989, 1991) on the estimated propensity score. We also compare with the BART model (Bayesian Additive Regression Trees; Chipman, et al. 2010), which provides a very flexible regression of observed outcomes on the treatment variable and the covariates (Hill, 2011). Extensive data-based simulation studies have shown that, in terms of bias and mean square error in causal effect estimation, these linear regression models and BART outperform normal linear regression of outcomes using (1) propensity-score-based pair-matching or subclassification alone, (2) treatment indicators and estimated propensity scores as covariates, and (3) observation weights defined by inverse of propensity score estimates, when the only covariate is a treatment indicator (Robins, et al. 2000), and when the linear model also includes subject covariates (Kang & Schafer, 2007; Schafer & Kang, 2008; Hill, 2011). These results seemed to hold true, especially when both the outcome and propensity score models were misspecified for the data, which, arguably, almost always occurs in practice.

Novelty of study:

The new BNP model is the first model that allows one to investigate how treatments causally effect the entire distribution (density) of the outcome variable, including any feature of the distribution that is of interest. Here we study the model for the analysis of observational data, though it can also be used to analyze data from a randomized study. Also, by comparing the models on predictive accuracy, we will answer an open question about whether the new BNP model can improve upon models that have been previously proposed for causal inference.

Statistical Model:

The BNP model is an infinite-mixture regression model, which allows the entire probability density (distribution) of the outcome variable to change flexibly with covariates; in the case of a discrete-valued outcome, the density is for the underlying latent response. The model consists of covariate-dependent mixture weights, defined by an ordered-probit regression that has an infinite sequence of random probit variances, and has systematic component defined by a random process, which we specify as the linear model with regression coefficients β. Each kernel of the infinite mixture is a possibly-distinct and general unimodal density, specifically, a scale-mixture of uniforms that is flexibly modeled by a nonparametric, stick-breaking prior (Ishwaran & James, 2001), which we specify as the Pitman-Yor process (more general than the DP). A feature of the BNP regression model is that, for any given covariate value, the outcome density becomes unimodal when the value is informative about the response, and becomes multimodal when it is not very informative. The BNP model is completed by assigning a prior distribution to all parameters of the model. For data analysis, the prior combines with the data via Bayes' theorem, to yield the posterior distribution, which describes the plausible values of all model parameters. This posterior distribution is estimated using Markov Chain Monte Carlo (MCMC) methods, including a Gibbs sampling method that is useful for infinite-mixture models (Kalli et al., 2010).

Research Setting, Subjects, Intervention, Design, and Data Collection:

Through the analysis of data arising from an observational (quasi-experimental) study, we investigate the causal effect of exposure to excellent high school math education (the treatment), versus non exposure (the control), on ACT math score (ACT, 2007), among 99 undergraduate teacher candidates who have recently started attending and learning at the education schools of Loyola University Chicago (LUC), Northeastern Illinois University (NEIU), National-Louis University (NLU), and the University of Illinois-Chicago (UIC). The four universities have partnered to address the need to improve K-12 math education at Chicago urban schools, and they obtained candidates' data from surveys, interviews, and admission reports. Of the 99 candidates, 25.3% of candidates said that they received excellent math teaching in high school. On average, ACT math scores were higher for those who were exposed to excellent teaching (22.16) versus non-excellent teaching (20.23). Table 1 presents summary statistics for 45 pretreatment (group-indicator 0-1) variables that describe each candidate's background.

Data Analysis Details and Plan:

For each member of a sample of teaching candidates indexed by $i = 1,\ldots,n$, let $\mathbf{x}_i = (x_{1i},\ldots,x_{pi})^{\mathrm{T}}$ denote a background (pre-treatment) covariate, let $T_i \in \{t = 0,1\}$ denote the treatment variable where $T_i = 1$ when a candidate is exposed to the treatment of excellent high school math instruction, and $T_i = 0$ when s/he received the control treatment of non-exposure, and let $(Y_i(1),Y_i(0))$ denote a candidate's potential ACT math outcomes in response to treatment and to

control. A causal effect is a comparison of potential outcomes, such as $Y_i(1) - Y_i(0)$. The fundamental problem of causal inference is that only outcome, namely $Y_i = t_i Y_i(1) + (1-t_i)Y_i(0)$, can be observed from each candidate, because s/he only receives one treatment, t_i (Holland, 1986). This makes causal effects not directly observable from the raw data.

However, in the observational study, causal effects can be identified if the sample data satisfy 3 assumptions (Imbens, 2004): (1) *The Stable-Unit Treatment Value Assumption (SUTVA)*: potential outcomes for one candidate are independent of potential treatment status of any another candidate, given the observed covariates; (2) *Unconfoundedness*: potential outcomes and treatment assignments are independent, given any value of **x**, and thus, given any true propensity score $e(\mathbf{x}) = \Pr[T = 1|\mathbf{x}]$ (Rosenbaum & Rubin, 1983); and (3) *Overlap*: any candidate with any given value of **x** has a chance to receive either treatment or control. Note that SUTVA can be weakened by introducing a multi-valued treatment variable that describes both the treatment received by the candidate and by the other candidates. Then the other two assumptions can be cast in terms of this treatment variable (Imai & van Dyk, 2004).

If all 3 assumptions hold, then $\mathrm{E}[Y(t) \mid \mathbf{x}] = \mathrm{E}[Y \mid T = t, \mathbf{x}] = \mathrm{E}[Y \mid T = t, e(\mathbf{x})]$ holds for any value of **x**, and for any $t = 0,1$, allowing causal effects to be identified by a regression model (Imbens, 2004), which preferably, admits consistent estimates of these conditional expectations.

From this perspective, the treatment variable T is simply another covariate in the regression model, and causal inference entails a comparison of regression predictions $\mathrm{E}[Y_i| T_i = 1, \mathbf{x}]$ and $\mathrm{E}[Y_i| T_i = 0,\mathbf{x}]$ for each candidate $i = 1,\ldots,n$, with the first prediction (second prediction, respectively) an "out-of-sample prediction" when $T_i = 0$ (when $T_i = 1$, respectively), as shown in the earlier research (Kang & Schafer, 2007; Schafer & Kang, 2008; Hill, 2011). For a given regression model, the accuracy of out-of-sample predictions can be assessed by leave-one-out cross-validated log-likelihood, CVLPL (Geisser & Eddy, 1979; Hastie, et al. 2009).

We use the CVLPL criterion to compare the predictive performance between 20 regression models, including the new BNP model, BART, and 18 linear regression models, most of which are based either on subclassification, full matching, direct regression, or inverse-weighting, by the estimated propensity score. This also includes two DP-mixed, Hierarchical Linear Models (HLMs), each a random ANCOVA model with candidates nested within subclasses, and fully-matched groups, respectively. For all 20 models, we assume that the data satisfy Assumptions 1-3, as is typically done in the practice of causal inference. Propensity scores were estimated from the data, by fitting a binary logit regression of the treatment variable T on **x**, using forward selection of 1035 variables, including main effects of the 45 background covariates, and all two-way interactions. Propensity score matching was done using the MatchIt (Ho et al., 2011) and optmatch (Hansen & Klopfer, 2006) packages of the R software (R Development Core Team, 2011). The BART model was fit to the data, using the BayesTree package of R (Chipman, and McCulloch, 2010), based on 150,000 converged MCMC samples. The 18 linear regression models were fit using the MATLAB software (2011, The MathWorks, Natick, MA). The DP-mixed HLM was fit using 50,000 converged MCMC samples, via the DP package (Jara, 2007).

More generally, when Assumptions 1-3 hold, $\mathrm{E}[\phi\{Y(t)\} \mid \mathbf{x}] = \mathrm{E}[\phi\{Y\}|T = t, \mathbf{x}] = \mathrm{E}[\phi\{Y\}|T = t, e(\mathbf{x})]$ holds for all values of **x**, for $t = 0, 1$, and for any choice of function, ϕ (Imbens, 2004). This includes not only the identity function $\phi(Y) = Y$, as implied earlier, but also includes other interesting choices of functions, such as the conditional distribution function, $\mathrm{E}[\phi\{Y(t)\}| \mathbf{x}] = \mathrm{E}[\mathbb{I}\{Y(t) < y\} | \mathbf{x}] = F(y | \mathbf{x})$ ($\mathbb{I}\{\cdot\}$ is the indicator function), and derivatives including the density function, quantiles, and the inter-quartile range (a robust measure of variance) (Imbens, 2004). Then for any choice of function, ϕ, the conditional average treatment effect is given by:

$$\text{CATE}_{\phi} = \frac{1}{n}\sum_{i=1}^{n} \text{E}[\phi\{Y(1)\} - \phi\{Y(0)\} \,|\, \mathbf{x}_i] = \frac{1}{n}\sum_{i=1}^{n} \{\text{E}[\phi\{Y\} \,|\, T = 1, \mathbf{x}_i] - \text{E}[\phi\{Y\} \,|\, T = 0, \mathbf{x}_i]\}. \quad (1)$$

This causal effect estimator avoids the questionable assumption that the sample is a random draw from a population (Imbens, 2004). The usual CATE estimator, $\text{CATE} = \frac{1}{n}\sum_{i=1}^{n} \text{E}[Y(1) - Y(0) \,|\, \mathbf{x}_i]$, assumes $\phi(Y) = Y$. Though, clearly, other choices of functions ϕ can be used to compare outcomes in terms of quantiles, inter-quartile range, and so on.

Inferences with these more complicated functionals can be easily obtained from the posterior distribution of our BNP regression model, via MCMC. Also, for the BNP model, we specified a weakly-informative proper prior distribution, which accurately reflects little prior information on the model parameters. After combining this prior with the data via Bayes' theorem, the resulting posterior of the model parameters will be based almost entirely on information from the sample data. The MCMC estimation of the model posterior distribution was performed using code we wrote in the MATLAB software, and was based on 50,000 converged MCMC samples. Also, for the BNP model, we define the covariates by indicator variables of 25 matched-groups of candidates, formed by optimal full matching, and interactions between the treatment indicator variable and these 25 matched-group indicator variables. This full-matching is based on absolute multivariate (L_1) distance between each pair of the 99 candidates on the covariate vector **x**, after replacing each coordinate in the vector with their empirical ranks (Rosenbaum, 1991). This matching was done using the optmatch package of R (Hansen & Klopfer, 2006).

Usefulness and Empirical Results of the Proposed Method (Model):

In the analysis of the observational data, Table 2 shows that, in terms of the CVLPL measure of predictive accuracy, our new BNP model far-outperforms the BART model, and all the 18 linear regression models mentioned earlier. All 20 models yielded positive estimates of the usual CATE estimator ($= \frac{1}{n}\sum_{i=1}^{n} \text{E}[Y(1) - Y(0) \,|\, \mathbf{x}_i]$), suggesting that exposure to excellent high school math education (versus non-exposure/control) causes an increase in ACT math scores. Most models concluded that the 95% predictive interval of CATE was significantly higher than 0.

From the posterior predictive distribution of the BNP model, Figure 1 and Table 3 present the estimated density, quantiles, and inter-quartile range of the ACT math scores, under exposure to excellent high math instruction, and under non-exposure (control), and the CATE density estimate. All these statistics were based on averaging over all the 99 candidates. Figure 1 also presents the posterior median and inter-quartile range of CATE, for each individual candidate.

Given these results, and given the fact that the new BNP regression model outperformed all other regression models that assumed normally-distributed errors, it appears that the outcome and causal effect distributions are truly non-normal, skewed, heavy-tailed, and either unimodal or multimodal. Also, the results suggest that exposure to excellent high school math instruction (vs. non-exposure/control) significantly increases ACT math achievement.

Conclusions:

Through the analysis of an observational data set on math achievement, we showed that the new BNP regression model can provide richer causal inferences with higher predictive accuracy, compared to typical causal models which focus inference on the mean outcome, and which make restrictive parametric assumptions about the outcome variable and about the propensity score model. The new BNP model allows one to investigate how treatments causally change any interesting aspect of the distribution (density) of (potential) outcomes, in a flexible manner.

Appendices

Appendix A. References

ACT. (2007). *The ACT technical manual*. Iowa City, IA: ACT, Inc.

Breslow, N., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9-25.

Chipman, H., and McCulloch, R. (2010). *BayesTree: Bayesian Methods for Tree Based Models*. http://CRAN.R-project.org/package=BayesTree.

Chipman, H., George, E., & McCulloch, R. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics, 4,* 266-298.

Geisser, S., & Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74,* 153-160.

Hansen, B., & Klopfer, S. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15,* 609-627.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd edition)*. New York: Springer-Verlag.

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20,* 217-240.

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

Ho, D.E., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric Preprocessing for parametric causal inference. *Journal of Statistical Software*. Forthcoming.

Imai, K., & Dyk, D. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99,* 854-866.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86,* 4-29.

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

Jara, A. (2007). Applied Bayesian Non- and Semi-parametric Inference using DPpackage. *Rnews, 7,* 17-26.

Kalli, M., Griffin, J., & Walker, S. (2010). Slice sampling mixture models. *Statistics and Computing, 21,* 93-105.

Kang, J., & Schafer, J. (2007). A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22,* 523-539.

Karabatsos, G., & Walker, S. (2011). *Bayesian unimodal density regression (Tech. Rep.)*. Chicago: University of Illinois.

Kleinman, K., & Ibrahim, J. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics, 54,* 921-38.

Kleinman, K., & Ibrahim, J. (1998b). A semiparametric Bayesian approach to generalized linear mixed models. *Statistics In Medicine, 17,* 2579-2596.

Laird, N., & Ware, J. (1982). Random effects models for longitudinal data. *Biometrics, 38,* 963-974.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models (Second Ed.)*. London: Chapman and Hall.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*.

R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
Robins, J., Hernán, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550-560.
Rosenbaum, P. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024-1032.
Rosenbaum, P. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society*, *Series B*, *53*, 597-610.
Rosenbaum, P., & Rubin, D. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.
Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516-524.
Schafer, J., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279-313.

Appendix B. Tables and Figures

Variable	%	Variable	%
AfAm	13.1	HS_PerfArts	3.0
AmInd	1.0	HS_Alternative	2.0
Asian	8.1	HS_Size<200	10.1
Latino	47.5	HS_Size 200-500	2.0
White	27.3	HS_Size 500-1k	11.1
Mixed	2.0	HS_Size 1k-1.5k	3.0
Female	87.9	HS_Size 1.5k-2k	18.2
Age 18	51.5	HS_Size 2.5k-3k	5.1
Age 19	38.4	HS_Size 3k-3.5k	11.1
Age 20	4.0	HS_Size>4k	7.1
Age 21	2.0	HS_LowInc 0-20%	20.2
Age 22-26	3.0	HS_LowInc 20-40%	14.1
Age ≥27	1.0	HS_LowInc 40-60%	11.1
HS_IL	94.9	HS_LowInc 60-80%	10.1
HS_public	87.9	HS_LowInc 80-100%	29.3
HS_private	11.1	HS_AfAm 60-90%	4.0
HS_urban	56.6	HS_AfAm 90-100%	2.0
HS_suburban	40.4	HS_Latino 60-90%	19.2
HS_rural	2.0	HS_Latino 90-100%	6.1
HS_cps	45.5	HS_mixed	35.4
HS_Select	3.0	HS_White 60-90%	19.2
HS_Magnet	1.0	HS_White 90-100%	1.0
HS_CollegePrep	16.2		

Table 1. Descriptive statistics of pre-treatment variables among the 99 teacher candidates, including variables of candidate high school background (labeled HS).

CVLPL	Out?	CATE	2.5%	97.5%	Regression (causal) model: Predictors
-177	N	1.53	.95	2.16	New BNP: ~I, FMrank, T by FMrank
-259	N	1.89	.29	3.48	HLM/ANCOVA/DP: I, T, 7pc(**x**) by FM
-259	N	1.88	.13	3.59	HLM/ANCOVA/DP: I, T, 7pc(**x**) by SC
-260	Y	1.63	1.04	2.23	IPW/OLS: I, T, 7pc(**x**)
-260	N	1.91	1.32	2.51	ANCOVA/OLS: I, T, 7pc(**x**)
-262	N	1.15	-.16	2.49	BART: T, **x**
-265	N	.68	.10	1.27	OLS: I, T by SC, 7pc(**x**)
-268	N	2.03	1.41	2.65	OLS: ~I, T, C, T by 5pc(**x**), C by 5pc(**x**)
-268	Y	1.78	1.16	2.41	IPW/OLS: I, T, T by 5pc(**x**), C by 5pc(**x**)
-269	Y	2.29	1.74	2.85	OLS: I, T by FM, 7pc(**x**)
-271	N	.33	-.35	1.01	OLS: I, T by SC, T by 1pc(**x**), C by 1pc(**x**)
-281	Y	1.38	.71	2.04	OLS: I, T by FM, T by 1pc(**x**), C by 1pc(**x**)
-286	N	.53	-.25	1.32	OLS: ~I, T,C, T by SC, C by SC
-287	N	.70	-.12	1.52	OLS: I, T by SC
-287	N	1.56	.70	2.42	IPW/OLS: I, T
-288	N	.77	-.09	1.62	OLS: I, T, ê(**x**) (polynomial, degree 1).
-288	N	.96	.10	1.82	OLS: I, T, logit(ê(**x**)) (polynomial, degree 1).
-291	Y	2.38	1.59	3.17	OLS: I, T by FM
-324	Y	1.81	1.10	2.52	OLS: ~I, T, C, T by FM, T by C
-419	Y	1.33	.83	1.83	OLS: I, T, **x**

Notes:

(1) Out?: Indicates whether any standardized residual indicated any outliers under the model.

(2) 2.5%, 97.5%: the 95% predictive interval bounds of CATE.

(3) New BNP: The new Bayesian nonparametric regression model, proposed in this paper.

(4) OLS: Linear regression (causal) model fit under ordinary least squares.

(5) HLM/ANCOVA/DP: A Dirichlet process mixed Hierarchical Linear Model, with a random ANCOVA model for each subclassified or fully-matched group.

(6) IPW: regression with each observation weighted by the estimated propensity score $(t_i/\hat{e}(\mathbf{x}_i)) + (1-t_i)/(1-\hat{e}(\mathbf{x}_i))$.

(7) Npc(x): N principal components of **x**, for dimension reduction, and to ensure positive-definiteness for OLS. For each model, the number of components (e.g., 7pc(x)) was chosen to maximize CVLPL.

(8) I: intercept; ~I: intercept excluded; T: treatment indicator; C: control indicator;

(9) **x** : vector of 45 variables describing a candidates background (see Table 1).

(10) ê(**x**): Estimated propensity score from a fitted binary logit regression.

(11) For the two regression (causal) models having ê(x) as a predictor, polynomials of ê(x) up to order 10 were considered. In each case, order 1 (linear) was found to maximize CVLPL.

(12) FM: 15 group indicators from full optimal matching of candidates on ê(**x**), into 16 groups.

(13) SC: 5 group indicators, from subclassification of candidates on ê(**x**), into 6 groups.

(14) FMrank: 24 group indicators, from full optimal matching of candidates on the ranking of the coordinates of **x**, into 25 groups.

(15) by: refers to an interaction, e.g., "T by SC".

Table 2. Predictive accuracy (CVLPL) and CATE estimates for 20 regression (causal) models.

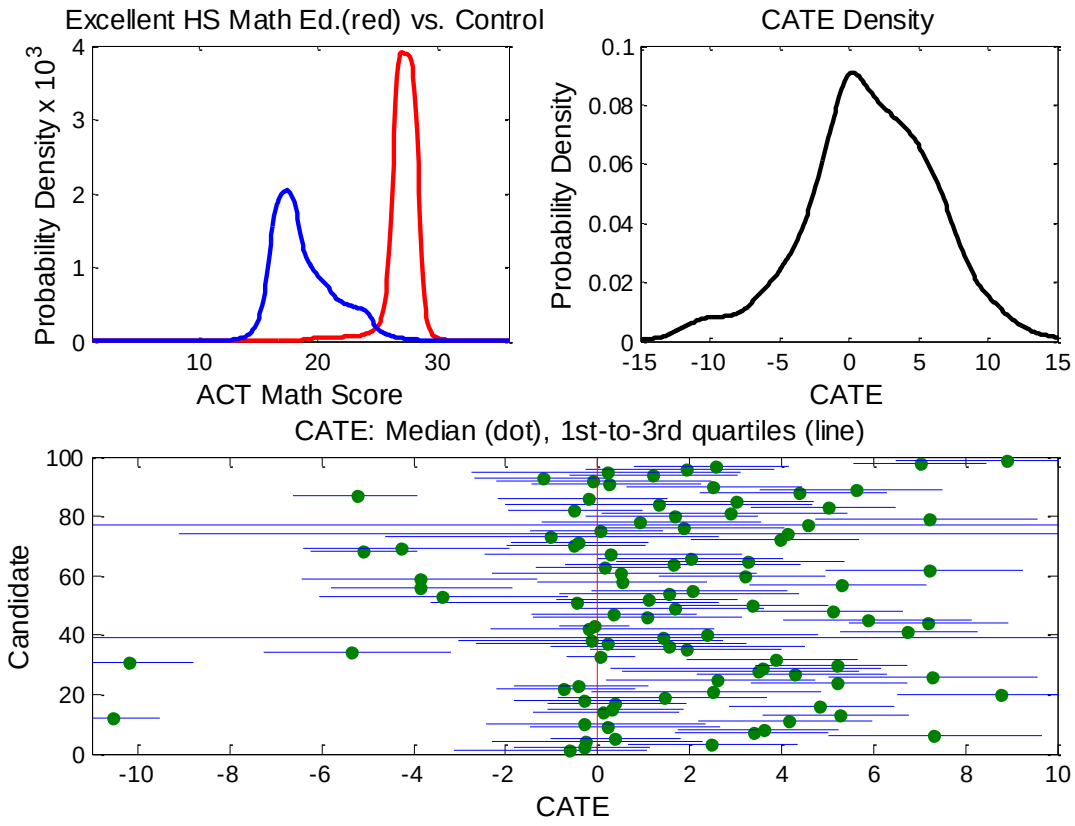

Figure 1. Top left panel: Posterior density estimated of ACT math scores, under the treatment (excellent high school math education), and under the control, averaged over the 99 candidates. Top right panel: CATE density posterior estimate, averaged over the candidates. Bottom panel: posterior median (dot) and inter-quartile range (line) of CATE, for each candidate.

ACT Math Score	10%	25%	50%	75%	90%	IQR
Under Excellent HS Math Education	17.81	20.23	21.96	23.56	25.7	3.33
Under Non-excellent (control)	17.65	18.89	20.31	21.71	23	2.81

Table 3. Posterior estimates of quantiles (percentiles) and inter-quartile range (IQR) of ACT math scores, under treatment and under control, averaged over candidates, and after controlling for candidate covariates.