NISS

Estimation of Propensity Scores Using Generalized Additive Models

Mi-Ja Woo, Jerome Reiter and Alan F. Karr

Technical Report Number 167
June 2007

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Estimation of Propensity Scores Using Generalized Additive Models

Mi-Ja Woo*, Jerome P. Reiter†, Alan F. Karr‡

Abstract

Propensity score matching is often used in observational studies to create treatment and control groups with similar distributions of observed covariates. Typically, propensity scores are estimated using logistic regressions that assume linearity between the logistic link and the predictors. When the actual assignment mechanism is not governed by linearity, matching on the poorly estimated propensity scores might not produce groups with similar covariate distributions. In this paper, we evaluate the use of generalized additive models (GAMs), which use flexible rather than linear functions of the predictors, for estimating propensity scores. Using empirical studies, we compare GAMs to logistic regressions in terms of balancing covariate distributions when matching on estimated propensity scores. We find that, when the distributions of covariates in the treatment and control groups overlap sufficiently, using GAMs can improve overall covariate balance, especially for higher order moments and fine features of distributions. When the distributions in the two groups overlap insufficiently, GAMs more clearly reveals this fact than logistic regression do.

Key words: Causal inference; Logistic regression; Observational study.

1 Introduction

In many studies, the objective is to learn about the causal effect of some treatment relative to a control condition. Ideally, the treatment and control condition

*National Institute of Statistical Sciences, Research Triangle Park, NC, USA.
†Duke University, Durham, NC, USA.
‡National Institute of Statistical Sciences, Research Triangle Park, NC, USA.

are randomly assigned to subjects. With large samples, random assignment produces groups with similar background characteristics prior to receipt of treatment, so that significant differences in the outcome variable in the two groups primarily reflect the effects of the treatment, not the effects of differences in covariates. When random assignment is not feasible, e.g., in observational studies, there is no guarantee that background characteristics are similar in the treatment and control groups. When these covariates are related to the outcome, any observed differences in the two groups' outcome distributions may reflect the differences in the groups' backgrounds rather than effects of the treatment [1, 2].

Analysts can reduce the bias that results from imbalanced covariate distributions, at least for observed covariates, using propensity score matching [3, 4, 5, 6]. The propensity score, $e(\mathbf{x})$, is the probability of being assigned to receive the treatment given a vector of covariate values $\mathbf{x}$. That is, $e(\mathbf{x}) = P(R = 1|\mathbf{x})$, where $R = 1$ if assigned to treatment and $R = 0$ if assigned to control. When two large groups have the same distributions of $e(\mathbf{x})$, the groups should have similar distributions of $\mathbf{x}$ [3]. Thus, to balance observed covariate distributions in observational studies, analysts can create a matched control group by selecting control records with similar propensity scores as the treated records. Analysts can check the success of the matching by verifying that the distributions of the covariates are similar in the two groups. If balance is satisfactory, the outcomes in this matched control group then are compared to the outcomes in the treatment group to estimate causal effects.

Propensity scores are rarely known exactly and must be estimated from the data. A commonly used approach is to fit a logistic regression of R on $\mathbf{x}$ with main effects for each causally-relevant $\mathbf{x}$, and use the estimated probabilities as the propensity scores. While easy to fit, this logistic regression assumes linearity between the logistic link function and the covariates. However, linearity might inaccurately describe the relationships between assignment and the covariates, which could impact the success of matching. Including high order polynomial functions of $\mathbf{x}$ can help account for non-linearity, but these can be unwieldy when there are many predictors in $\mathbf{x}$, particularly when interactions are included in the logistic regression.

In this paper, we investigate the usefulness of generalized additive models [7], abbreviated as GAMs, for estimating propensity scores. GAMs replace the linear component of logistic regression with a flexible additive function. Because this forces fewer restrictions on the propensity score model, using GAMs may improve balance after matching. As done in other studies of propensity score methodology [8, 9, 10, 11], we evaluate this hypothesis using empirical studies. In Section 2, we

simulate assignment to treatment given two covariates taken from genuine data, and perform propensity score matching based on logistic regressions and GAMs. The studies show that using GAMs can improve covariate balance over using logistic regressions, particularly for higher order moments and fine features of the covariate distributions. In Section 3, we construct two observational studies—one with adequate overlap in covariate distributions and one with minimal overlap—using genuine data. When there is sufficient overlap, the results again favor GAM over logistic regression. When there is insufficient overlap, the GAM reveals this problem more clearly than the logistic regression does. Finally, in Section 4, we provide some general remarks about estimation of propensity scores.

2 Empirical Study Using Simulated Assignments

The covariate data, $\mathbf{x}$, comprise 5,000 heads of households and two variables, person age (x_A) and household property taxes (x_P), randomly sampled from the March 2000 Current Population Survey. We simulate treatment assignments by drawing from Bernoulli random variables with $e(\mathbf{x}) = Pr(R = 1|\mathbf{x})$ computed from four different scenarios:

$$\log\left(\frac{e(\mathbf{z})}{1+e(\mathbf{z})}\right) = -1.6 - .7z_A - .8z_P \quad (1)$$

$$\log\left(\frac{e(\mathbf{z})}{1+e(\mathbf{z})}\right) = -.6 - .8z_A^2 - .7z_P^2 - .9z_Az_P \quad (2)$$

$$\log\left(\frac{e(\mathbf{z})}{1+e(\mathbf{z})}\right) = -1.5 + .6z_A^3 - .7z_P^3 - .4z_Az_P^2 \quad (3)$$

$$\log\left(\frac{e(\mathbf{z})}{1+e(\mathbf{z})}\right) = -.5 - .7z_A^{.5} - .4z_Az_P^2 - .2\frac{z_P}{z_A} - .3z_A^4 - .6z_P^6 \quad (4)$$

where $\mathbf{z} = (z_A, z_P)$ is a standardized version of $\mathbf{x} = (x_A, x_P)$ computed with the means and variances of all 5,000 records. We refer to the scenario in (1) as the linear case, in (2) as the quadratic case, in (3) as the cubic case, and in (4) as the complex case. For each scenario, around 4,000 units are assigned to the control group and 1,000 to the treated group. We simulate 100 allocations of treatment assignments.

We consider propensity score estimation using four models. All models include x_A and x_P as predictors. These four models are (i) a logistic regression with main effects only, (ii) a logistic regression with main effects and interactions, (iii)

a GAM with main effects only, and (iv) a GAM with main effects and interactions. We use four degrees freedom for the smoothing splines in the GAMs. After estimating the propensity scores, $\hat{e}(\mathbf{x})$, for each treated unit j we find the unit k among all control records such that $|\hat{e}(\mathbf{x}_j) - \hat{e}(\mathbf{x}_k)|$ is minimized. For simplicity, we match with replacement.

To evaluate balance in the covariate distributions after matching, in each replication i, where $1 \leq i \leq 100$, we compute the means, standard deviations, and the tenth and ninetieth percentiles of x_A and x_P for both the treated and matched control groups. We also compute the correlation, ρ, between x_A and x_P in each group. For each of these quantities, we compute the mean squared difference across the 100 replications between the treatment and matched control group quantities,

$$MSD_k = \sum_{i=1}^{100} (q_{ik}^{(t)} - q_{ik}^{(mc)})^2/100, \tag{5}$$

where $q_{ik}^{(t)}$ and $q_{ik}^{(mc)}$ represent the value of the kth quantity in replication i for the treated and matched control groups, respectively. Values of $MSD_k = 0$ indicate that the treated and matched control units have exactly the same value of the kth quantity in all 100 replications. For any one quantity, the value of MSD_k is relatively large when the treated and control units values of that quantity are usually far apart in the replications.

Table 1 displays the MSDs for the four assignment scenarios and four propensity score models. It also displays the MSDs when using the true propensity score functions in (1) - (4) to estimate $e(\mathbf{x})$. The MSDs associated with property tax quantities are much larger than those associated with age quantities because property tax ranges from 0 to 100,000, whereas age ranges from 15 to 90. Property tax has many zero values, which explains why $MSD = 0$ for the tenth percentile of property tax.

For the linear case, all four models perform reasonably well at balancing covariates. In aggregate, the GAM with interactions slightly outperforms the others, as it has the smallest or second smallest MSD across the quantities, including noticeably smaller values for the standard deviations and the correlation. The competitive performance of the logistic regression models is not surprising here, since linearity is a correct assumption. However, once linearity does not apply, the GAMs do a much better job at balancing covariate distributions, particularly for standard deviations: the MSDs are typically orders of magnitude smaller when using the GAMs. Including interactions in the models improves balance in the correlations, although this sometimes is accompanied with degraded balance on

means and percentiles. We note that matching on estimated propensity scores typically performs better than matching on true ones, as has been recognized by [12] and [13].

3 Constructed Observational Studies

We now construct two observational studies using data from a randomized experiment and from the National Longitudinal Survey of Youth (NLSY). These data also were used by [14] and [15], who illustrated the benefits of propensity score matching using logistic regressions to estimate $e(\mathbf{x})$. The experiment is the Infant Health and Development Program (IHDP), which randomized 377 low birth weight infants to an intensive child care program and 608 low birth weight infants to a control program. Details on the design of the experiment are available in [16], [17], and [18].

As described in [15], the NLSY is a panel survey that began in 1979 with a sample of approximately 12,000 teenagers who, appropriately weighted, were nationally representative at that time. These participants were interviewed every year thereafter until 1994 and biannually after that. We restrict the NLSY data to the 4,511 children born from 1981 to 1989, because the IHDP began in 1985. We assume that none of the NLSY children received the intensive child care. Thus, the reservoir of control units in our constructed observational study comprises the 4,511 NLSY and 608 IHDP infants assigned to the control group. We simplify analyses by using only records with completely observed data, which leaves 357 treated units and 4263 control units.

The background variables include mother's ethnicity, mother's age and educational attainment at time of birth of child, whether worked in the year before the child was born, child's sex, birth weight, age in months, and indicator for born pre-term. As evident in Table 2, these covariates are not balanced in the treated and full control groups. The distributions of birth weight, days in the hospital, weeks pre-term, and mothers' races are very different in the two groups. Table 3 shows that some correlations also are far apart, notably the correlation between mother's and child's age and weeks pre-term and child's age.

We therefore use propensity score matching to create groups of matched controls. Because the 608 controls from the IHDP are in the full control reservoir, we know that the covariate distributions overlap adequately. As in Section 2, we estimate propensity scores using logistic regressions and GAMs, both with main effects only in the models. The match for each treated child is the child in the full

control reservoir whose estimated propensity score is closest to the treated child's score. We report the results from matching with replacement; matching without replacement produces very similar covariate distributions.

As evident in Tables 2 and 3, propensity score matching improves covariate balance, regardless of whether one uses logistic regression or GAM. The two models do not produce radically different balance in terms of means. Still, when comparing the two models, the means of the matched control group based on the GAM are closer to the means for the treated group for 10 out of 17 covariates. The story for higher order moments is clearer. Matching with GAM generates better balance with respect to the standard deviations, particularly for birth weight and child's age. And, as evident in Table 3, matching with GAM is more effective at balancing the correlations.

The box plots in Figure 1 show further evidence of the advantages of GAM over logistic regression in these data. Age and birth weight have very different distributions in the treated and full control groups. While matching with logistic regression improves balance, the distributions of these two covariates remain noticeably different from those in the treatment group even after matching. After matching based on GAM, the distributions more faithfully resemble those in the treated group.

Because the IHDP controls were in the reservoir, we were guaranteed to have decent overlap in the covariate distributions. We now examine a case without this guarantee. Specifically, we drop the 608 controls from the reservoir, leaving only the NLSY infants available for matching. As shown in Figure 2, the propensity scores for the full control group differ dramatically from those for the treatment group when using the GAM. We would be reluctant to match with such disparity. Indeed, as displayed in Table 4, matching with replacement using GAM performs poorly; for example, all matched controls received prenatal care and none are Hispanic. In fact, because of the severe lack of overlap in the propensity scores, one control observation is used as a match 301 times. Matching without replacement does not fix the lack of balance, leading to many matched control children with higher birth weights and fewer days in the hospital than treated children.

The propensity score distributions for the treated and full control groups overlap more when using logistic regression, as shown in Figure 3, than when using the GAM. After matching, this results in better covariate balance with respect to means and standard deviations, as shown in Table 4. However, substantial imbalance remains in the distributions of birth weight (the means differ by more than two standard deviations), child age (much higher variability for matched controls), mother's age (smaller variability for matched controls), and education levels.

In this dataset with minimal overlap, the GAM sharply discriminates between the treated and control records, whereas the logistic regression does not. This is not a failure of the GAM; rather, it reflects differences in features of the covariate distributions that are not picked up by the linear function of the predictors in the logistic regression.

4 Concluding Remarks

The simulations here suggest that estimating propensity scores with generalized additive models can improve balance in covariates relative to estimating propensity scores with logistic regressions, at least when there is adequate overlap in the covariate distributions. GAMs eliminate the need to include many high order terms, such as quadratic or cubic terms, which can be advantageous when there are many potentially relevant confounders to be included in the models.

Even with the evidence presented here, we do not recommend that analysts rule out logistic regression as a viable approach to estimating propensity scores. Analysts can perform the matching using both models, and select the matched control group that gives the best overall balance. Balance should be judged not only by similarity of means, but by similarity of entire distributions, which can be examined with histograms and other graphical tools.

This article did not discuss methods of matching other than nearest neighbor. However, we did try an alternative algorithm, genetic matching [19], in the constructed observational study. When the covariate distributions had minimal overlap, using genetic matching produced matched control groups with similar characteristics whether using GAM or logistic regression to estimate propensity scores. This suggests that the matching method and possibly its interaction with the estimation method can substantially impact covariate balance. Future study is needed to investigate this interaction.

References

[1] Cochran WG. The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A* 1965; **128**:134–155.

[2] Rubin DB. Estimating casual effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.

[3] Rosenbaum PR, Rubin, DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.

[4] Rubin DB. Estimating casual effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**:757–763.

[5] D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281.

[6] Stuart EA, Rubin, DB. Matching methods for causal inference: designing observational studies. In *Best Practices in Quantitative Social Science*. Osborne J (ed). Sage Publications: Thousand Oaks, CA, in press.

[7] Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman and Hall: Boca Raton; 1990.

[8] Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**:1231–1236.

[9] Huppler Hullsiek K, Louis T. Propensity score modeling strategies for the causal analysis of observational studies. *Biostatistics* 2002; **2**:179–193.

[10] Austin PC. A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2006; **25**:2084–2106.

[11] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* 2007; **26**:734–753.

[12] Thomas N, Rubin DB. Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 1996; **52**:249–264.

[13] Joffe MM, Rosenbaum PR. Invited commentary: Propensity scores." *American Journal of Epidemiology* 1999; **150**(4):327–333.

[14] Hill J, Reiter JP, Zanutto E. A comparison of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Gelman, A., Meng, X.L. (eds). Wiley: New York, 2004; 49–60.

[15] Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine* 2006; **25**:2230–2256.

[16] Infant Health and Development Program. Enhancing the outcomes of low-birth-weight, premature infants. *Journal of the American Medical Association* 1990; **22**:3035–3042.

[17] Brooks-Gunn J, Liaw FR, Klebanov PK. Effects of early intervention on cognitive function of low birth weight preterm infants. *Journal of Pediatrics* 1991; **120**:350–359.

[18] Hill J, Brooks-Gunn J, Waldfogel J. Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology* 2003; **39**:730–744.

[19] Diamond A, Sekhon J. Genetic matching for estimating casual effects: a general multivariate matching method for achieving balance in observational studies. http://sekhon.berkeley.edu/papers/GenMatch.pdf 2005.

	AGE				PROPERTY TAX				
	Mean	SD	10%	90%	Mean	SD	10%	90%	ρ
Linear Case									
Logistic with Main	.093	.044	.215	1.105	1390	1821	0	11488	.00087
Logistic with Inter	.088	.048	.218	.891	1327	2135	0	11483	.00086
GAM with Main	.085	.054	.204	1.100	1438	1978	0	8367	.00095
GAM with Inter	.066	.039	.152	1.002	1400	1537	0	8670	.00052
True $e(\mathbf{x})$	.054	.064	.174	.071	850	3670	0	9620	.00104
Quadratic Case									
Logistic with Main	3.10	9.646	4.203	52.94	1529	28732	0	16468	.012
Logistic with Inter	.82	8.568	6.102	39.13	2141	233657	0	17462	.002
GAM with Main	.29	.012	.273	.95	1640	920	0	10400	.017
GAM	.11	.021	.341	.52	2330	1533	0	13669	.0006
True $e(\mathbf{x})$	.15	.027	.287	.70	1327	2277	0	15730	.0011
Cubic Case									
Logistic with Main	1.46	1.56	.42	7.95	17971	10775	0	89730	.013
Logistic with Inter	1.74	2.17	.37	10.50	19074	9775	0	90982	.019
GAM with Main	.15	.16	.23	1.18	12141	5317	0	104658	.017
GAM	.16	.25	.27	1.80	11322	5241	0	85633	.013
True $e(\mathbf{x})$	.15	.21	.20	1.38	9896	6662	0	110845	.018
Complex Case									
Logistic with Main	1.80	10.87	3.77	48.48	1588	41337	0	24993	.00257
Logistic with Inter	2.46	12.20	3.31	58.93	3765	278315	0	25346	.00614
GAM with Main	.10	.099	.25	.54	1853	1297	0	16163	.00654
GAM	.13	.014	.43	.81	2805	1448	0	17914	.00069
True $e(\mathbf{x})$	.15	.015	.17	.66	5601	4537	0	59402	.00098

Table 1: Mean squared differences between treated and matched control groups for covariate distributions.

	Treated	Full Control	Matched Control	
	Mean (SD)	Mean (SD)	Logistic	GAM
Mother				
Age(yrs)	24.8 (5.8)	23.8 (3.7)	24.4 (5.5)	24.5 (5.7)
Hispanic	.11 (.31)	.19 (.39)	.07 (.26)	.09 (.29)
Black	.53 (.40)	.32 (.46)	.55 (.50)	.52 (.50)
White	.36 (.48)	.49 (.50)	.37 (.48)	.38 (.49)
Married	.43 (.49)	.66 (.47)	.48 (.50)	.43 (.49)
No HS	.41 (.49)	.31 (.46)	.40 (.49)	.39 (.49)
HS	.29 (.45)	.41 (.49)	.30 (.46)	.31 (.46)
Some college	.17 (.38)	.20 (.40)	.20 (.40)	.18 (.39)
College	.13 (.33)	.08 (.28)	.10 (.30)	.12 (.33)
Working	.60 (.49)	.62(.48)	.60 (.49)	.61 (.49)
Prenatal care	.95 (.21)	.98 (.12)	.95 (.21)	.97 (.18)
Child				
Birth Weight	1821 (435)	3106 (785)	1783 (529)	1809 (428)
Days in Hospital	23.6 (22.7)	7.4 (13.5)	24.1 (21.6)	24.9 (24.8)
Age 1990 (mos.)	56.8 (2.0)	57.8 (27.2)	57.4 (15.4)	56.7 (2.0)
Weeks Pre-term	6.9 (2.5)	2.0 (3.0)	7.0 (2.8)	7.1 (2.7)
Sex (1=female)	.51 (.50)	.50 (.50)	.48 (.50)	.50 (.50)
First born	.47 (.50)	.43 (.49)	.43 (.49)	.48 (.50)

Table 2: Means and standard deviations of the covariates in treated, full control, and matched control groups, including the 608 controls from the IHDP.

		Age	Birth Weight	Days in Hospital	Weeks Pre-term
Mom Age	Treated	.069	-.051	.049	.076
	Full Control	-.567	-.039	.097	.098
	Logistic	-.193	-.058	.115	.131
	GAM	.089	-.050	.117	.086
Age	Treated		-.102	.184	.147
	Full Control		-.032	.007	-.036
	Logistic		.054	.004	.021
	GAM		-.221	.263	.241
Birth Weight	Treated			-.800	-.753
	Full Control			-.625	-.731
	Logistic			-.833	-.669
	GAM			-.808	-.756
Days in Hospital	Treated				.753
	Full Control				.681
	Logistic				.778
	GAM				.735

Table 3: Correlations of the covariates in the treated, full control, and matched control groups, including the 608 IHDP controls.

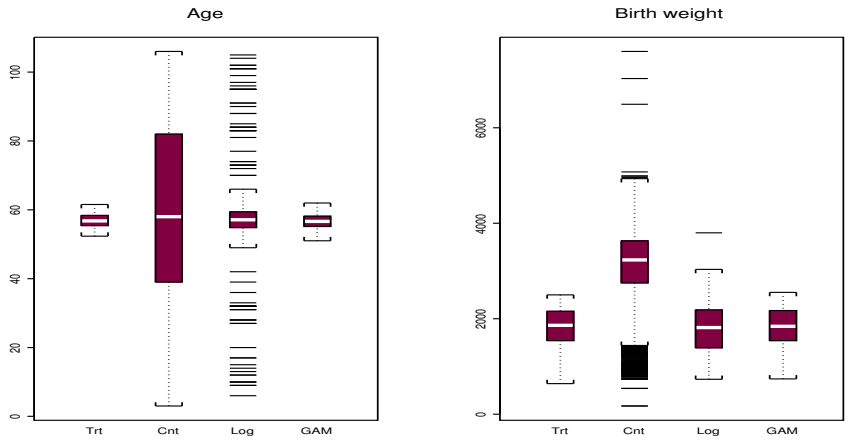

Figure 1: Box plots of age and birth weight in the constructed observational study, including the 608 controls from the IHDP.

	Treated	Full Control	Matched	
	Mean (SD)	Mean (SD)	Logistic	GAM
Mother				
Age(yrs)	24.8 (5.8)	23.6 (3.2)	24.0 (3.2)	24.0 (1.2)
Hispanic	.11 (.30)	.21 (.40)	.15 (.36)	.00 (.00)
Black	.53 (.50)	.29 (.45)	.49 (.50)	.85 (.35)
White	.36 (.48)	.51 (.50)	.36 (.48)	.15 (.35)
Married	.43 (.49)	.69 (.46)	.48 (.50)	.11 (.31)
No HS	.41 (.49)	.30 (.46)	.36 (.48)	.06 (.23)
HS	.29 (.45)	.43 (.49)	.42 (.49)	.10 (.30)
Some college	.17 (.38)	.19 (.39)	.17 (.38)	.003 (.05)
College	.13 (.33)	.08 (.27)	.05 (.22)	.84 (.36)
Working	.60 (.49)	.63(.48)	.64 (.48)	.90 (.29)
Prenatal care	.95 (.21)	.99 (.11)	.99 (.12)	1.00 (.00)
Child				
Birth Weight	1821 (435)	3314 (599)	1688 (591)	1086 (431)
Days in Hospital	23.6 (22.7)	4.5 (7.3)	25.7 (26.3)	60.6 (22.4)
Age 1990 (mos.)	56.8 (2.0)	57.9 (29.2)	54.4 (26.3)	58.7 (.9)
Weeks Pre-term	6.9 (2.5)	1.2 (2.1)	7.1 (3.8)	10.2 (2.1)
Sex (1=female)	.51 (.50)	.51 (.50)	.52 (.50)	.05 (.21)
First born	.47 (.50)	.43 (.49)	.47 (.50)	.88 (.33)

Table 4: Means and standard deviations of covariates in treated, control and matched groups, excluding the 608 controls from the IHDP.

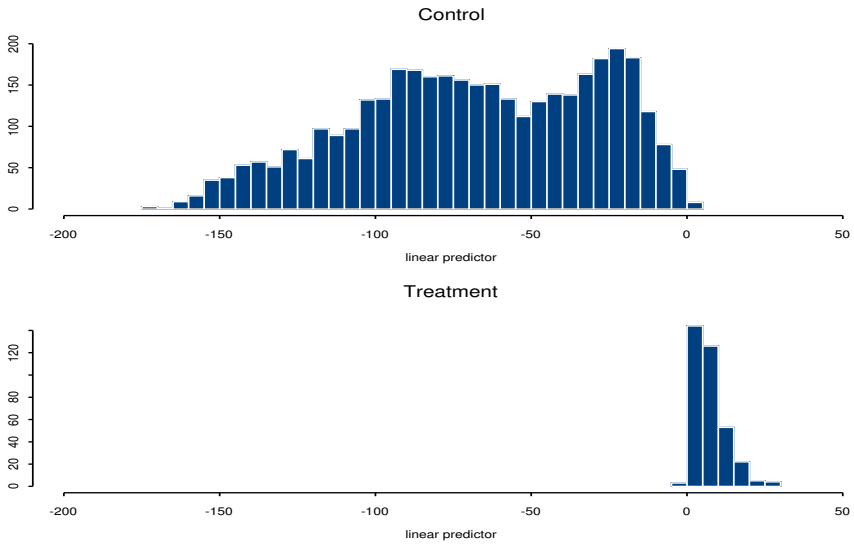

Figure 2: Histograms of the linear predictors for the propensity score when using the GAM, excluding the 608 controls from the IHDP

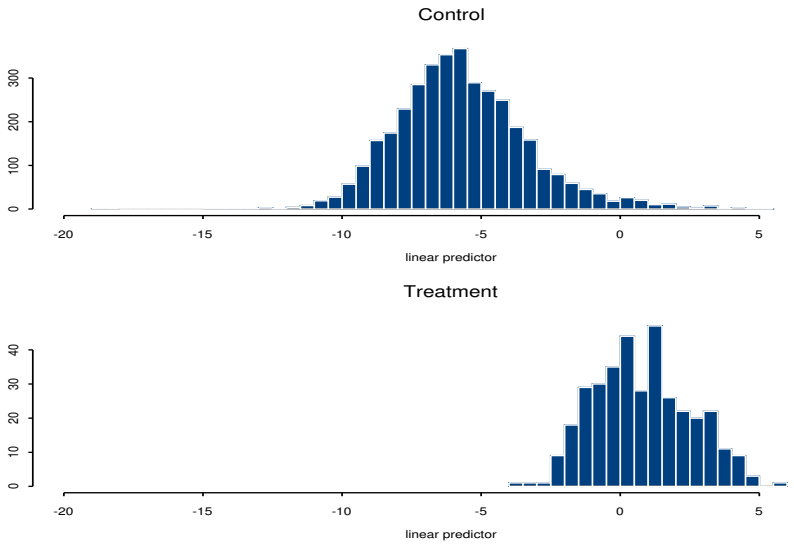

Figure 3: Histograms of the linear predictors for the propensity score when using the logistic regression, excluding the 608 controls from the IHDP