Biometrics 000, 000–000 DOI: 000

000 0000

Regression Adjustment and Stratification by Propensity Score in Treatment Effect Estimation

Jessica A. Myers* and Thomas A. Louis**

Department of Biostatistics, Bloomberg School of Public Health

Johns Hopkins University, Baltimore, Maryland, U.S.A.

**email:* jamyers@jhsph.edu

***email:* tlouis@jhsph.edu

Summary: Propensity score adjustment of effect estimates in observational studies of treatment is a common technique used to control for bias in treatment assignment. In situations where matching on propensity score is not possible or desirable, regression adjustment and stratification are two options. Regression adjustment is used most often and can be highly efficient, but it can lead to biased results when model assumptions are violated. Validity of the stratification approach depends on fewer model assumptions, but is less efficient than regression adjustment when the regression assumptions hold. To investigate these issues, by simulation we compare stratification and regression adjustments. We consider two stratification approaches; equal frequency classes and an approach the attempts to minimize the mean squared error (MSE) of the treatment effect estimate. The regression approach we consider is a Generalized Additive Model (GAM), that flexibly estimates the relations among propensity score, treatment assignment, and outcome. We find that, under a wide range of plausible data generating distributions, the GAM approach outperforms stratification in treatment effect estimation with respect to bias, variance, and thereby MSE. We illustrate approaches via analysis of data on insurance plan choice and its relation to satisfaction with asthma care.

Key words: Generalized Additive Model; Observational study; Optimal stratification; Propensity score adjustment.

1. Introduction

In observational studies where investigators seek to estimate the effect of a binary treatment ("treatment" and control), treatment assignment is not randomized. As a result, treatment groups may differ substantially on potentially confounding covariates, biasing estimated treatment effects (Rubin, 1991; Sommer and Zeger, 1991). Methods available to control for confounding include regression, matching, and stratification on covariates (Cochran, 1968; Billewicz, 1965). Propensity score methods, developed by Rosenbaum and Rubin (1983), may also be used to balance the distribution of measured covariates across treatment groups.

The propensity score of an experimental unit is the conditional probability of assignment to the treatment group, given observed covariates. Under complete randomization, this probability is controlled by the investigator and is stochastically independent of covariates. When units are not randomized, the propensity is induced by the assignment process. Specifically, units (individuals) that are treated will tend to have higher propensity scores than those who go untreated. This imbalance in propensity score represents an imbalance on covariates between treatment and control groups. Methods of adjustment utilizing the propensity, including matching, stratification, and regression adjustment on propensity scores, have been shown to yield unbiased estimates of treatment effect when the estimand of interest is the expected difference in response between treatment and control and treatment assignment is 'strongly ignorable' (Rosenbaum and Rubin, 1983, 1984, 1985; Dehejia and Wahba, 2002).

Matching can be highly effective in removing imbalance in covariates between treatment groups, but there are often some study units that cannot be matched and must be left out of analysis (D'Agostino Jr., 1998). The treatment effect estimate will then be based on a reduced, and potentially non-representative, set of cases. Thus, if investigators wish to estimate average treatment effect for the entire study population, regression adjustment and stratification are both useful options, but no consensus exists as to which method is preferable under various conditions.

The utility of each method of adjustment will depend on the mean squared error (MSE) of the resulting treatment effect estimate with lower MSE preferred.

When using the stratification approach, the range of propensity scores is split into strata, and treatment effect is estimated within each stratum. Then, overall treatment effect is computed by a weighted mean of the stratum-specific estimates. Stratification on propensity score does not require specification of the propensity-outcome relation, and, therefore, may be preferable to regression adjustment, especially when this relation is believed to be complex. However, choice of the number and placement of strata does influence the variance and bias of the combined estimate. Generally, there are opposing effects; wide strata produce low variance but high potential bias, narrow strata the reverse.

The most common implementation of stratification on propensity score is five equal frequency strata. A result from Cochran (1968), cited in Rosenbaum and Rubin (1983), indicates that approximately 90% of the initial bias due to the propensity is eliminated by this stratification. Importantly, Cochran's result is based on a linear relation between propensity and outcome. In other situations, such as those where stratification on propensity score is most desirable, stratification on the quintiles, for example, may not adequately remove bias, and other approaches to forming strata may be preferable. Hullsiek and Louis (2002) propose choosing strata that balance the variances of the stratum-specific estimates. This method generally produces an effect estimate with lower variance than the equal frequency approach because equal frequency strata with very high estimated variances will be widened to achieve variance balance. Of course, this widening can increase bias.

Reviews of propensity score methods in published clinical research have found that regression adjustment by propensity score is the most commonly used method, although investigators often fail to check the adequacy of their model specification (Shah et al., 2005; Weitzen et al., 2004). When the relation between propensity and outcome is linear, direct adjustment may be achieved by including a linear term for propensity in the regression model. In this scenario, regression adjustment is preferable to stratification because it estimates treatment effect with lower variance than stratification and similarly removes nearly all the bias (Rosenbaum and Rubin, 1984, 1983; D'Agostino Jr., 1998). When the relation between propensity and outcome is not linear, regression adjustment by propensity score will require more care. Specifying the propensity-outcome relation in a suitably flexible way, for example using Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1990), may control bias with a smaller variance than that for stratification.

We present a Monte Carlo study comparing the performance of regression adjustment and stratification approaches with respect to variance, bias, and MSE for several data generating models. The methods considered include equal frequency stratification on propensity score, an 'optimal' stratification on propensity score that minimizes the estimated MSE of the resulting treatment effect estimate, and regression adjustment on propensity score using GAMs. We assume throughout that the propensity score has been estimated well; however, deviations from this assumption would certainly effect the performance for all of the methods investigated. Section 2 describes notation and the propensity score methods under consideration. Section 3 presents the simulation study and results. Section 4 presents an analysis of an observational study of the effect of health insurance type on satisfaction with asthma care. Section 5 summarizes our findings.

2. Model and Methods

Let Z_i indicate treatment assignment, with $Z_i = 1$ for treatment and $Z_i = 0$ for control. Define the response vectors accordingly, $\boldsymbol{Y}_z = (Y_{1z}, Y_{2z}, \ldots, Y_{n_z z})$, where n_z is the sample size for treatment group z. Furthermore, let $\boldsymbol{X}_i$ be a vector of potential confounders, associated with both treatment and outcome. The propensity score is defined $e_i = e(\boldsymbol{X}_i) = Pr(Z = 1|\boldsymbol{X}_i)$. We assume linear confounding via the model

$$\begin{aligned} Y|z,e &= \beta_0 + \beta_1 z + g(e) + \epsilon \qquad (1) \\ \epsilon &\sim N(0,\sigma^2) \end{aligned}$$

where β_0 and β_1 are scalar parameters and g is some smooth function. Our target of estimation is the true average treatment effect, given by $\Delta = \beta_1$.

We are interested in comparing estimation approaches for Δ with respect to MSE and its components, variance and bias, which may be summarized:

$$MSE(\hat{\Delta}) = Var(\hat{\Delta}) + Bias(\hat{\Delta})^2 \qquad (2)$$

With no confounding, a simple difference of means, $(\bar{Y}_1 - \bar{Y}_0)$, is minimum variance, unbiased (and therefore minimum MSE) for estimating the treatment effect. In the presence of confounding, this estimate is biased. The initial bias for this unadjusted estimate is

$$\begin{aligned} Bias(\bar{Y}_1 - \bar{Y}_0) &= E(\bar{Y}_1) - E(\bar{Y}_0) - \beta_1 \\ &= \int_0^1 g(u)[f_1(u) - f_0(u)]du \qquad (3) \end{aligned}$$

where f_1 and f_0 are the densities of propensity scores in the treatment and control groups, respectively. We consider regression and stratification on propensity score for reducing bias in estimation of Δ.

2.1 *Regression adjustment*

The assumed model in (1) suggests the use of GAMs for estimating treatment effect. Fitting the GAM, $E(Y|z,e) = \beta_0 + \beta_1 z + g(e)$, treatment effect and variance are returned as the estimate and variance of the coefficient on treatment, β_1. The smooth term for propensity score, $g(e)$, is approximated as a sum of spline terms. Any of the well-known basis functions may be used in this sum; we use thin plate regression splines with cross-validated smoothing parameter selection, as described in Wood (2003, 2004). This regression model should closely mirror the true data generating process under our assumed model. If data are generated from another model, for example, a model with a non-additive effect of treatment, then the GAM will not represent the data generating process, but may be used for estimation of average treatment effect.

2.2 *Stratification*

Let $\boldsymbol{t} = (0 = t_0 < t_1 < \ldots < t_K = 1)$ define a partition of the range of propensity scores with K subclasses. Within each stratum, $k \in \{1, \ldots, K\}$, treatment effect is estimated with a simple difference of means, $\hat{\Delta}_k$. The variance of the difference of means is estimated

$$V_k \;=\; \hat{\sigma}^2_{1k}/n_{1k} + \hat{\sigma}^2_{0k}/n_{0k} \tag{4}$$

where $\hat{\sigma}_{zk} = \hat{V}ar(Y|Z = z, t_{k-1} < e \leq t_k)$ and n_{zk} is the number of units in treatment group z and subclass k. The overall treatment effect estimate and its estimated variance is the inverse-variance-weighted mean of the subclass-specific estimates, given by

$$\begin{aligned} \hat{\Delta} &= \left(\sum_{k=1}^{K} \hat{\Delta}_k V_k^{-1}\right) / \left(\sum_{k=1}^{K} V_k^{-1}\right) \\ \hat{Var}(\hat{\Delta}) &= 1/\left(\sum_{k=1}^{K} V_k^{-1}\right). \end{aligned} \tag{5}$$

Several authors have noted that the variance estimator given here generally underestimates the variance of the stratified treatment effect estimate because it treats the partition as fixed, rather than data dependent (Tu and Zhou, 2002; D'Agostino Jr., 1998). We use this estimator nonetheless because we are not primarily interested in the performance of the variance estimator, and we focus instead on the performance of the treatment effect estimators.

We consider two methods for choosing $\boldsymbol{t}$. In equal frequency stratification, the partition is defined by the quantiles. As an alternative, we seek to choose a partition that minimizes the MSE of the combined treatment effect estimate. In order to find the optimal partition, we must be able to estimate the MSE of the inverse variance-weighted estimator of treatment effect for a given partition. The estimator for variance is given above in (5). Under the assumed model (1), a formula for bias in stratum k is given by:

$$\begin{aligned} Bias(\hat{\Delta}_k) &= E(Y|Z=1, e \in (t_{k-1}, t_k)) - E(Y|Z=0, e \in (t_{k-1}, t_k)) - \beta_1 \\ &= E(g(e)|Z=1, e \in (t_{k-1}, t_k)) - E(g(e)|Z=0, e \in (t_{k-1}, t_k)) \\ &= \int_{t_{k-1}}^{t_k} g(x)[f_1(x)/M_{1k} - f_0(x)/M_{0k}]dx \qquad (6) \end{aligned}$$

where $M_{zk} = \int_{t_{k-1}}^{t_k} f_z(x)dx$. This formula can be estimated using the estimated functional form of the relation between propensity score and outcome, $\hat{g}$, returned by the GAM described above. In addition, we estimate the densities of propensity scores in each treatment group, f_1 and f_0, using a simple kernel density estimator.

Overall estimated bias of the treatment effect estimator is the inverse-variance-weighted mean of the subclass-specific biases. Using the estimates for bias and variance, we produce a function that returns the estimated MSE for a given partition, $\boldsymbol{t}$, and dataset, $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{e})$. The optimal partition for K subclasses is then found by treating the $K-1$ elements of $\boldsymbol{t}$ between 0 and 1 as the variable parameters in an optimization algorithm for minimizing the estimated MSE function.

3. Simulation Study

As shown in (3) and (6), the amount of bias due to propensity relates to the amount of imbalance in propensity between the two groups. We consider a class of conditional densities of the propensity scores in the treatment and control groups, defined respectively by:

$$\begin{aligned} f_1(e) &= \begin{cases} (2e)^s & e \leq .5 \\ 2-(2(1-e))^2 & e > .5 \end{cases} \\ f_0(e) &= f_1(1-e) \end{aligned}$$

These densities are by construction anti-symmetric ($f_z(e) = f_{1-z}(1-e)$) and produce a marginal uniform density when there is equal sample size in the two groups. Varying s produces a wide range of plausible distributions; an s of zero indicates a uniform distribution in each treatment group, and a large s indicates an extreme imbalance in propensity score by treatment assignment. We set $\pi = P(Z = 1) = 1/2$ and generated $N = n_1 + n_0$ propensity scores from a uniform distribution on (0,1). We then assigned treatment indicators according to

$$Z_i \sim Bernoulli(\pi f_1(e_i)). \tag{7}$$

We consider outcomes generated by two different models: the additive model, which is equivalent to the assumed model presented in (1), and the non-additive model,

$$\begin{aligned} Y_i &= exp\{\beta_0 + \beta_1 z_i + g(e_i) + \epsilon_i\} \qquad (8) \\ \epsilon_i &\sim N(0, \sigma^2). \end{aligned}$$

where, as before, β_0 and β_1 are scalar parameters and g is a smooth function. For both models, we consider the treatment effect of interest to be the average difference in expected outcome between treatment and control, conditional on propensity score. In the additive model, this quantity is given by $\Delta = \Delta(e) = \beta_1$. In the non-additive model, this quantity is equal to

$$\begin{aligned} \Delta &= \int_0^1 \Delta(e) dF(e) \\ &= \int_0^1 E(Y|Z = 1, e) - E(Y|Z = 0, e) dF(e) \\ &= exp\{\beta_0 + \sigma^2/2\}(exp\{\beta_1\} - 1) \int_0^1 exp\{g(e)\} dF(e) \qquad (9) \end{aligned}$$

3.1 *Simulation Settings*

We considered samples of size $N = 200$ and, as mentioned earlier, set $\pi = 0.5$ to achieve approximately equal sample size in the two groups and a marginal uniform distribution for propensity scores. We generated data with three different values of s: $.5$, 1, and 2, corresponding to low, moderate, and high imbalance, respectively, in propensity scores in the two groups. Furthermore, for each value of s, four different functional forms for g, the relation between propensity and outcome, are considered: (A) $g(e) = e$, (B) $g(e) = e^2$, (C) $g(e) = 2(e - .3)^2$, and (D) $g(e) = .5 + 4(e - .5)^3$. Each of these functions has an approximate range of [0,1] over the same domain interval, and, therefore, each induces a similar amount of bias. Relation (A) is linear, relation (B) is non-linear, but monotonic, relation (C) is non-linear and non-monotonic, and relation (D) has a non-monotonic first derivative. The three values for s and four functions for g yield 12 simulation scenarios for each of the two models.

In the additive model, we chose $\beta_0 = 0$, $\beta_1 = .25$, and $\sigma = .5$, so that the true treatment effect is equal to one half of the error standard deviation. In the non-additive model, we chose $\beta_0 = -.125$, $\beta_1 = .2$, and $\sigma = .5$. Because we specified $\beta_0 = -\sigma^2/2$ and $e \sim Unif(0,1)$, the expression for the treatment effect in the non-additive model is reduced to $(exp\{\beta_1\} - 1) \int_0^1 exp\{g(u)\}du$. We simulated 1000 datasets under each scenario and each model.

With each simulated dataset, we applied each of the methods presented in Section 2, including: (1) adjustment by propensity score via a smooth term for propensity score in a GAM; (2) equal frequency (EF) stratification using $K = 1$ through 6 subclasses; and (3) optimal stratification, using $K = 1$ through 6 subclasses, chosen such that the estimated MSE of the resulting estimate is minimized. Additionally, in the simulations generated from the additive model, we use the true $g(e)$ in a generalized linear model (GLM) to compare with the GAM. In the GLM, the true values of $g(e_i)$ enter the model as an offset, so that only the intercept and treatment effect must be estimated. Clearly, this model cannot be estimated in practice, since one generally won't know the true propensity-outcome relation. In this study, we compare the estimates from this model to that obtained using GAM to show the amount of bias and variance in the GAM estimates that is due to estimation of the relation $g(e)$.

3.2 *Simulation Results*

In all of the data-generating scenarios considered, significant positive bias exists when treatment effect is estimated directly, corresponding to stratification with $K = 1$. Initial bias is similar (but not constant) across the four propensity-outcome relations because the ranges of these functions are similar. Varying the amount of imbalance in propensities, indexed by s, varies the amount of initial bias. Data generated under higher s values produce estimates of treatment effect with higher initial bias. Data simulated with high imbalance in propensity scores between treatment groups also suffer from lack of sufficient overlap; when using the stratification approaches, particularly with datasets simulated under $s = 1$ or $s = 2$, the outermost strata contain data from only one treatment group. Therefore, no treatment effect estimate is possible for those strata, and their corresponding variance estimates are infinite. In those situations, we allowed the infinite variance to dictate a zero weight for the data in those strata, so that the number of strata actually used in treatment effect estimation, denoted by K^*, is smaller than K, the number of strata intended. In this section, we present a selection of the simulation results, but results for all simulations discussed are available in Web Supplement A.

Figure 1 shows the average estimated treatment effect with one observed standard error bars (left panel), observed standard errors and average estimated standard errors with 95% quantile bars (center panel), and observed root MSE (right panel) for data simulated under the additive model with linear relation between propensity and outcome (relation (A)). Data is displayed for simulations using all three values of s and for all analysis approaches considered. The horizontal axis is K, the number of strata used, where $K = 0$ refers to the use of non-stratification methods, GLM and GAM. The use of $K = 1$ means no stratification (direct estimation through a simple difference of means); these estimates show the amount of initial bias. For $K > 2$, the number of simulations out of 1000 that have $K^* = K$ is printed above the corresponding plotting point for EF stratifications, and below the corresponding point for optimal stratifications.

The treatment effect estimate plots show that the GAM and both stratification approaches are effective at reducing or eliminating bias due to propensity score. In particular, for each value of s, the GAM produces estimates of treatment effect that are on average unbiased and nearly identical to that of the GLM. Estimating the propensity-outcome relation in the GAM increases the standard error of the treatment effect estimator only slightly compared to the GLM. Bias reduction through stratification is achieved better at larger values of K. The optimal stratification method does slightly outperform equal frequency stratification with respect to bias reduction at moderate values of K, but at large values of K, the bias is equivalent for both stratifications (or even slightly favoring EF stratification) and the observed standard error is smaller for EF stratification.

The standard error plots in Figure 1 confirm that our variance estimator for the stratified treatment effects does on average produce estimates of standard error lower than what is observed across simulations. Observed standard errors generally increase as K increases, although for $s = 2$, this is not the case because so many of these datasets had $K^* < K$. The standard error estimates resulting from the GAMs is on average close to the observed standard error and generally lower than the observed standard errors resulting from stratification.

The plots of root MSE (RMSE) in Figure 1 show that in these data the GAM results in lower RMSE than the stratification approaches, regardless of the value of s. The differences in RMSE between the GLM, GAM, and stratification approaches become larger as s increases. In addition, RMSE is approximately constant for stratification approaches with $K \geq 3$. Although we continue to reduce bias as we increase K, this bias reduction is paid for with increasing variance, thus leaving RMSE essentially constant.

[Figure 1 about here.]

In Figure 2, we display the same information as that plotted in Figure 1, except for $s = .5$ only and for propensity-outcome relation (B). The patterns are primarily the same as they were when the propensity-outcome relation was linear. GAM again provides an unbiased estimate with equivalent or smaller observed standard error than either stratification method. The differences in RMSE among estimation methods are again larger at larger values of s. The simulation results for relations (C) and (D) are available in the Web Supplement and are similar to the results presented here.

[Figure 2 about here.]

Figure 3 displays simulation results for data simulated under the non-additive model with $s = .5$ and propensity-outcome relation (A). There are several differences in these results compared to the results plotted in Figure 1 for the additive model. First, the true treatment effect is no longer equal to $\beta_1 = 0.25$, but is given by (9). Although the GAM estimates the true treatment effect well, the stratified estimates appear to be converging to some negatively biased quantity as K increases. These stratified estimates do pass through the truth at $K = 3$, but of course, a priori the analyst has no way of knowing which K to choose to achieve these results.

The insufficient overlap in the optimal strata seems to be amplified in the non-additive data compared to the additive data. In Figure 3, nearly all of the simulated datasets have $K^* = K$ when using EF stratification, but many datasets have $K^* < K$ when using optimal stratification. Also, the GAM is now generally overestimating the standard error of the estimates of treatment effect compared to that which is observed. In general, the GAM again has lower RMSE than either stratification method, regardless of the number of strata used.

[Figure 3 about here.]

In Figure 4, results are shown, as in Figure 3, for the simulations with propensity-outcome relation (B). Results are similar to those with propensity-outcome relation (A) and again show that the GAM outperforms stratification. The results for other values of s and the other propensity-outcome relations (C) and (D) are similar.

[Figure 4 about here.]

3.3 *Generality of Distributional Assumptions*

Above, we simulated propensity scores with a Uniform(0,1) marginal density and conditional distributions that are anti-symmetric. The former assumption is made without loss of generality because propensity scores not satisfying this condition may be transformed. If the conditional densities are anti-symmetric and we have equal sample size in each group, the uniform transform does not corrupt this property; however, if anti-symmetry is not present, it cannot be forced through a monotone transform.

Let $F(e)$, $F_1(e)$, and $F_0(e)$ be the cumulative distribution functions of propensity scores, marginally, in the treated group, and in the control group, respectively. Using F as the uniform transform, the transformed scores and their conditional distributions are given by

$$\begin{aligned} U &= F(e) \\ F^*_z(u) &= F_z(F^{-1}(u)) \end{aligned}$$

The conditional distribution of the transformed data is closely related to the conditional distribution of the untransformed data. Because of this relation, anti-symmetry is preserved under this transformation, as shown in the Appendix

The preservation of anti-symmetry under the uniformity transform follows from the more general fact that any monotone transform will preserve anti-symmetry. This property also implies that no monotone transform will produce anti-symmetry in data where it does not already exist. Therefore, the results presented above are at least partially generalizable to cases which do not meet the assumptions held thus far. It is possible that data without anti-symmetric conditional densities of propensity scores will produce different results. In some preliminary investigations of this possibility, results were very similar when propensity scores were simulated from densities that were not anti-symmetric.

4. Analysis of Insurance Plan Choice Data

The following analysis considers data collected on 2515 asthma patients as part of the 1998 Asthma Outcomes Survey (Masland et al., 2000). This study was initiated by the Pacific Business Group on Health and HealthNet health plan for the purpose of assessing the quality of asthma care from 20 physician groups. Huang et al. (2005) developed propensity score methods to address physician group as a multiple treatment analysis. Because we prefer a binary treatment, our analysis evaluates the effect of health insurance type on satisfaction with asthma care across the 20 providers. Insurance type is classified as public, purchased through an employer, purchased personally, or other. A large majority, 2360 individuals, held either employer or personally purchased health insurance, and we consider the subset of data with these two insurance types so that the treatment of interest is dichotomous. Our indicator of treatment, Z, indicates having personally purchased health insurance.

The outcome is also dichotomous; $Y = 1$ indicates very good or excellent satisfaction with care, and $Y = 0$ indicates less than very good satisfaction. We are interested in estimating the average difference in the probability of high satisfaction with care between individuals with personally purchased and employer purchased insurance plans, controlling for confounders of treatment assignment and outcome. Clearly, this example is different from the data simulated in the Monte Carlo studies because those data all had continuous outcomes. However, our goal of estimation here is the same as in the simulations, and we may expect that the estimation problems faced in data simulated from the non-additive model will be similar to the problems faced in these data. Therefore, we follow the suggestion of Hellevik (2008) and use a Gaussian family GAM, exactly as implemented in the simulations, to estimate treatment effect.

We began by considering the measured covariates available for use in the propensity score, which include information about demographics, medical care, and health status. Demographic covariates are age (18-56), race (Black, White, Asian/Pacific Islander, American Indian, Other), Hispanic identification, gender, educational attainment (high school or less, college, post-graduate work), and employment status (none, part-time, full-time). Covariates that describe subjects' medical care are primary physician specialty (pulmonary/allergy specialist, other), consistent care by the same provider, physician group (1-20), and drug insurance coverage. Health status covariates include smoking (none, moderate, high), physical activity in the last four weeks (1-7), severity of asthma (1-4), comorbidity count (0-8), number of years with asthma (1-54), and the SF36 Health Survey composite scores for physical and mental health (0-100).

We must choose, of the measured covariates listed above, which to include in the propensity score model. Studies of propensity score methods have found that best results are achieved by only including covariates that are associated with outcome (Austin et al., 2007; Brookhart et al., 2006). This selection should include all of the potential confounders, those covariates associated with both treatment assignment and outcome. Therefore, before we estimate any propensity score models, we check each covariate by fitting a logistic regression model of outcome on treatment and the covariate. These models allow us to determine if there is an association between covariate and outcome when controlling for treatment assignment and to order the covariates with respect to their effect on outcome, as recommended by Hill (2008). For nominal categorical covariates, we fit simple GLMs, and for continuous or ordinal categorical covariates, we fit GAMs. Checks of association for the 11 categorical covariates and the 6 continuous covariates, respectively, are displayed in Web Supplement B. From these figures, we determined that when adjusting for treatment only smoking, employment status, and physical activity seem to share no association with the outcome, satisfaction with asthma care.

In the spirit of flexible model estimation, we used a logistic GAM of the personal health insurance indicator on the remaining 14 covariates that are associated with outcome to estimate the propensity score for each individual (Woo et al., 2008). The propensity score obtained is the predicted probability of holding personally purchased insurance, rather than employer purchased insurance, given model covariates. We ran an all subset regression with the eight most important covariates (always present in the model) and some subset of the other six predictors. We compared the unbiased risk estimator (UBRE) of these 64 models to identify a smaller set of useful candidate models. For each candidate model, we then checked the balance of all 14 covariates associated with outcome to identify our final model for propensity score estimation. Balance was checked through side-by-side boxplots of covariates, stratified on both treatment and propensity score quintile, or through two-by-two tables of treatment and covariates within propensity score quintiles. Figures in Web Supplement B show the balance checks for the final model chosen, which included: (1) random intercepts for physician groups; (2) main effects for race, education, consistent provider care, drug coverage, years with asthma, physical composite score, and mental composite score; and (3) a smooth term for age, which we note has a nonlinear relation with the log odds of treatment. Older and younger adults are more likely to have personally purchased health insurance than adults in middle-age.

[Figure 5 about here.]

Figure 5 shows the densities of the propensity scores in both treatment groups. The two groups overlap well with respect to propensity score, and we can estimate average treatment effect for the entire propensity score range. We next apply each of the three methods considered in the Monte Carlo study: GAM estimation, EF stratification, and optimal stratification. In addition, we compare these propensity score-based methods with the usual regression of outcome on the covariates used in the propensity score model and the treatment indicator.

Figure 6 displays the treatment effect estimation results of all analysis approaches considered. All analyses estimate a statistically significant or nearly statistically significant positive effect of holding personally purchased health insurance on satisfaction with asthma care. In particular, the GAM with propensity score approach, estimates that, on average, the probability of being highly satisfied with asthma care is 0.047 (-0.004, 0.097) larger for individuals with personally purchased health insurance than individuals with employer purchased health insurance, controlling for propensity to treatment. This estimate is reduced slightly from the unadjusted treatment effect estimate, 0.061 (0.015, 0.107). In Web Appendix B, we show the estimated smooth term for propensity score as estimated by the GAM. There is a small positive relation between propensity score and outcome, and this association reflects the confounded relation between treatment and outcome.

[Figure 6 about here.]

5. Discussion

The objective of this study was to compare the relative merits of stratification and regression approaches utilizing the propensity score for estimating treatment effects in observational studies and to explore the potential of an 'optimal' stratification procedure. Stratification on propensity score and regression adjustment with a smooth term for propensity score in a GAM both estimate treatment effect flexibly, allowing for nonlinear association between propensity score and outcome, and both are effective at reducing bias due to propensity to treatment. Based on the results from the Monte Carlo simulations, we recommend the GAM approach for three reasons: the GAM generally produces estimates with lower average bias and variance; the GAM requires less user choice to achieve bias reduction compared to stratification where the analyst must, at minimum, choose an appropriate K; and, thanks to flexible and automated GAM packages, such as `mgcv` in R, the GAM is simpler to implement than even EF stratification. In addition, the lack of necessary user choice in GAMs allows the outcomes to stay "hidden" until the final step of analysis, as advocated by Rubin (2001, 2007).

The benefits of GAMs in this case do not, however, overcome the need for great care in propensity score analysis. For example, analysts must still check for covariate balance on estimated propensity score. In the analysis presented in Section 4, we checked approximate balance of covariates within propensity score quintile, which ensures unconfounding of treatment and outcome within quintile. How best to check for balance when the propensity score will be used in a covariate regression has not been studied. In the Monte Carlo studies, we assumed that we have a well-estimated propensity score, and that there exists a "true" smooth relation between the estimated propensity score and the expectation of outcome. In data where covariates are not well-balanced by the estimated propensity score, we may expect that these assumptions fail and the results for all methods considered will be worse than what is presented here.

Estimated propensity scores must also be checked for sufficient overlap of treatment groups. Insufficient overlap may result in a modified estimand or inappropriate extrapolation, regardless of the propensity score analysis method used. In particular, in each of the simulations presented in this paper, we additionally implemented a GAM that estimated a separate smooth term for propensity score among treated and untreated subjects. We then estimated average treatment effect using this model to predict the unobserved potential outcomes. We did not present the results from this method in Section 3.2 because the imbalance in the tails of the propensity score distributions led to inappropriate extrapolation and extremely poor estimates of average treatment effect. The GAM with a single smooth term for propensity score is partially protected from this kind of extrapolation because the estimated effect of treatment is forced to be constant across the range of propensity scores. Therefore, treatment effect is estimated primarily from data units that lie in overlapping regions of the propensity score distributions; however, this case may result in an estimand that is different than what the investigator intended.

Finally, we note that regression adjustment may be more problematic when variances differ between treatment groups. We investigated this possibility in the non-additive simulation studies, where data was simulated with heteroscedastic errors. The stratification approaches allow for differing variance estimates between treatment groups and across strata. The GAM approach does not model the heteroscedasticity, but still outperformed stratification in these simulations.

In cases of more extreme imbalance in variances between treatment groups, we may find that the GAM no longer performs well.

Supplementary Materials

Web Appendices and Figures referenced in Sections 3.2 and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgments

This research was supported by Grant 5T32ES012871 from the U.S. National Institute of Environmental Health Sciences and Grant R01 DK061662 from the U.S. National Institute of Diabetes, Digestive and Kidney Diseases. The authors wish to thank I-Chang Huang and Constantine Frangakis for supplying the data analyzed in this paper.

References

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). 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* **26,** 734–753.

Billewicz, W. (1965). The efficiency of matched samples: An emperical investigation. *Biometrics* **21,** 623–643.

Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J., and Sturmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology* **163,** 1149.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24,** 295–313.

D'Agostino Jr., R. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* **17,** 2265–2281.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics* **84,** 151–161.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Hellevik, O. (2008). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality and Quantity* **43,** 59–74.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'by Peter Austin, Statistics in. *Statistics in medicine* **27,** 2055–2061.

Huang, I., Frangakis, C., Dominici, F., Diette, G., and Wu, A. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research* **40,** 253–278.

Hullsiek, K. H. and Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics* **2,** 179–193.

Masland, M., Wu, A., Diette, G., Dominici, F., and Skinner, E. (2000). The 1998 asthma outcomes survey. *San Francisco, CA: Pacific Business Group on Health* .

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79,** 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39,** 33–38.

Rubin, D. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* **2,** 169–188.

Rubin, D. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine* **26,** 20.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **47,** 1213–1234.

Shah, B., Laupacis, A., Hux, J., and Austin, P. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of clinical epidemiology* **58,** 550–559.

Sommer, A. and Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10,** 45–52.

Tu, W. and Zhou, X. (2002). A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *Health Services and Outcomes Research Methodology* **3,** 135–147.

Weitzen, S., Lapane, K., Toledano, A., Hume, A., and Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety* **13,** 841–853.

Woo, M., Reiter, J., and Karr, A. (2008). Estimation of propensity scores using generalized additive models. *Statistics in medicine* **27,** 3805–3816.

Wood, S. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association* **99,** 673–687.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* **65,** 95–114.

Appendix Proof of Anti-symmetry Preservation

In this section we prove that when $\pi = 1/2$, anti-symmetry of conditional densities is preserved under the uniform transform. Recall, anti-symmetry is defined, $f_z(e) = f_{1-z}(1-e)$. Equivalently, we may state, $F_z(e) = 1 - F_{1-z}(1-e)$. We must show that $F_1^*(1-u) = 1 - F_0^*(u)$.

First, note that

$$\begin{aligned}
F(1-e) &= \frac{1}{2}F_1(1-e) + \frac{1}{2}F_0(1-e) \\
&= \frac{1}{2}[1 - F_0(e)] + \frac{1}{2}[1 - F_1(e)] \\
&= 1 - \frac{1}{2}F_1(e) - \frac{1}{2}F_0(e) \\
&= 1 - F(e) \\
\Rightarrow \quad F(1 - F^{-1}(e)) &= 1 - e \\
\Rightarrow \quad 1 - F^{-1}(e) &= F^{-1}(1-e).
\end{aligned}$$

Then consider

$$\begin{aligned}
F_1^*(1-u) &= F_1(F^{-1}(1-u)) \\
&= F_1(1 - F^{-1}(u)) \\
&= 1 - F_0(F^{-1}(u)) \\
&= 1 - F_0^*(u).
\end{aligned}$$

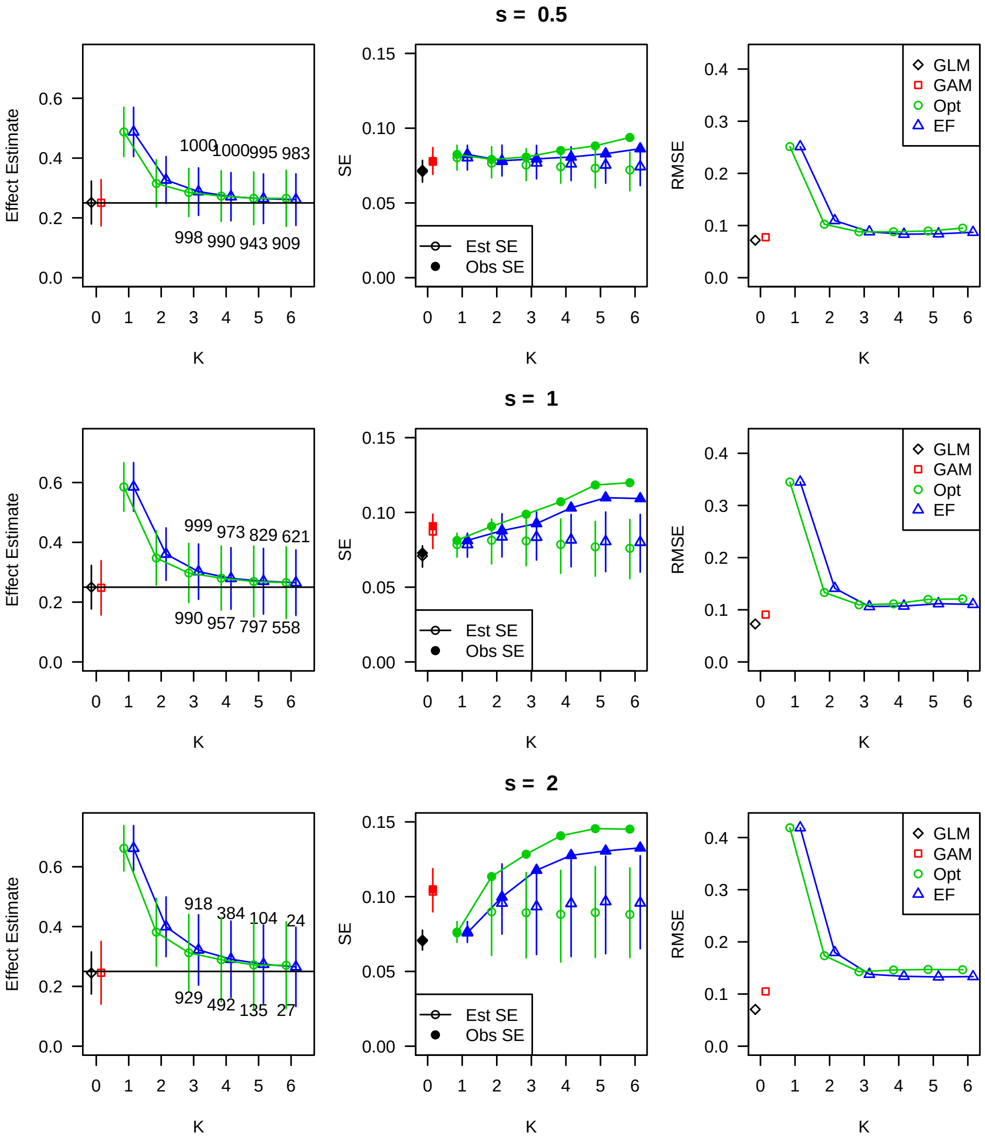

Figure 1. Average estimated treatment effect with one observed standard error bars (left panel), observed standard errors and average estimated standard errors with 95% quantile bars (center panel), and observed root MSE (right panel) for data simulated under the additive model with linear relation between propensity and outcome (relation (A)). Data is displayed for simulations using all three values of s and for all analysis approaches considered. The horizontal axis is K, the number of strata used, where $K = 0$ refers to the use of non-stratification methods, GLM and GAM, and $K = 1$ means no stratification (direct estimation through a simple difference of means).

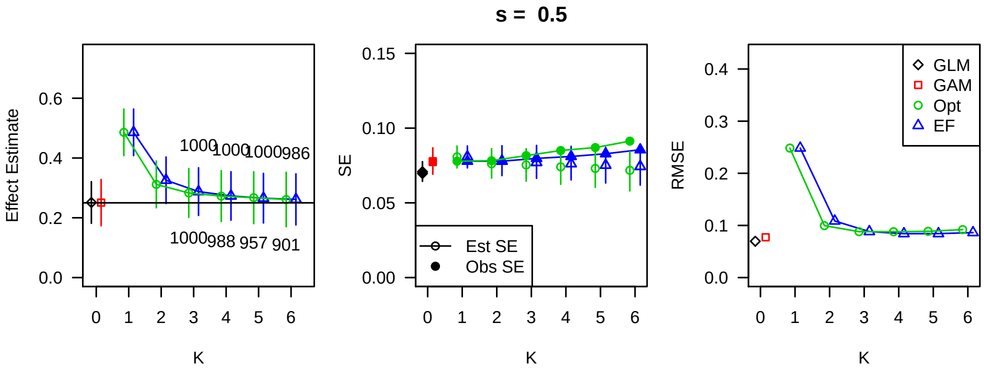

Figure 2. Simulation results for data simulated under the additive model with $s = .5$ and propensity-outcome relation (B), corresponding to $g(e) = e^2$.

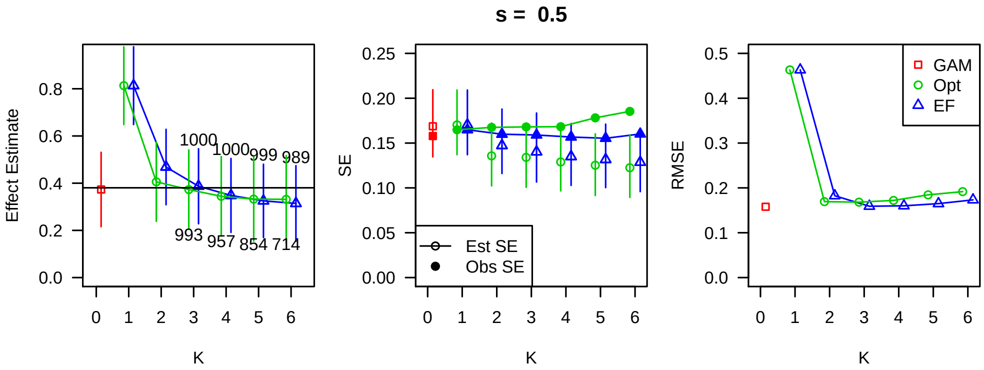

Figure 3. Simulation results for data simulated under the non-additive model with $s = .5$ and propensity-outcome relation (A), corresponding to $g(e) = e$.

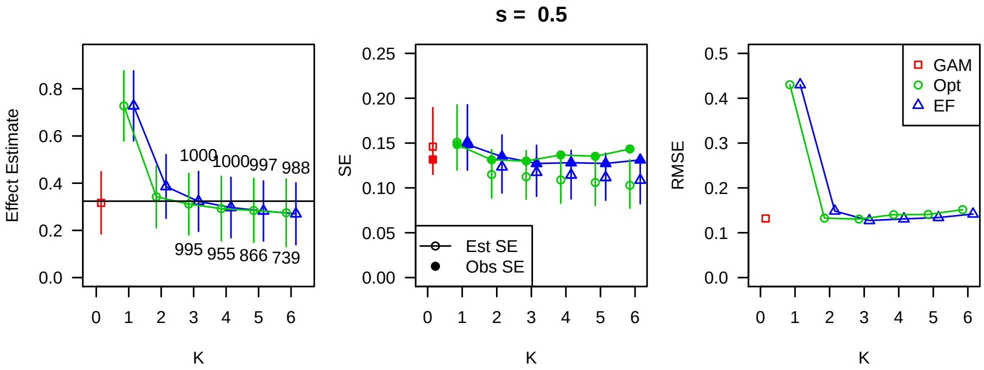

Figure 4. Simulation results for data simulated under the non-additive model with $s = .5$ and propensity-outcome relation (B), corresponding to $g(e) = e^2$.

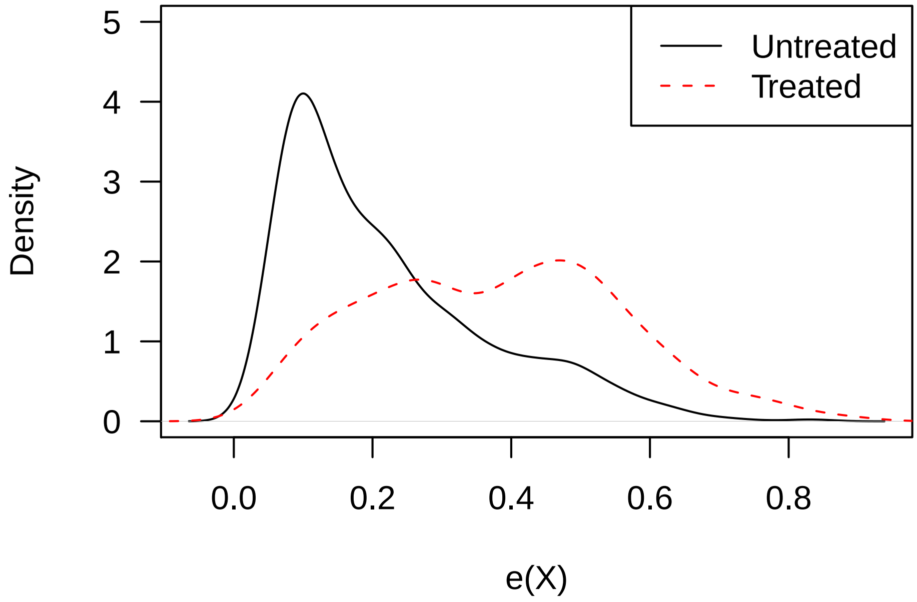

Figure 5. Relative frequencies of the estimated propensity scores conditional on treatment. Only 26.4% of units had personally purchased health insurance ("Treated"), and 73.6% of units had employer purchased health insurance ("Untreated").

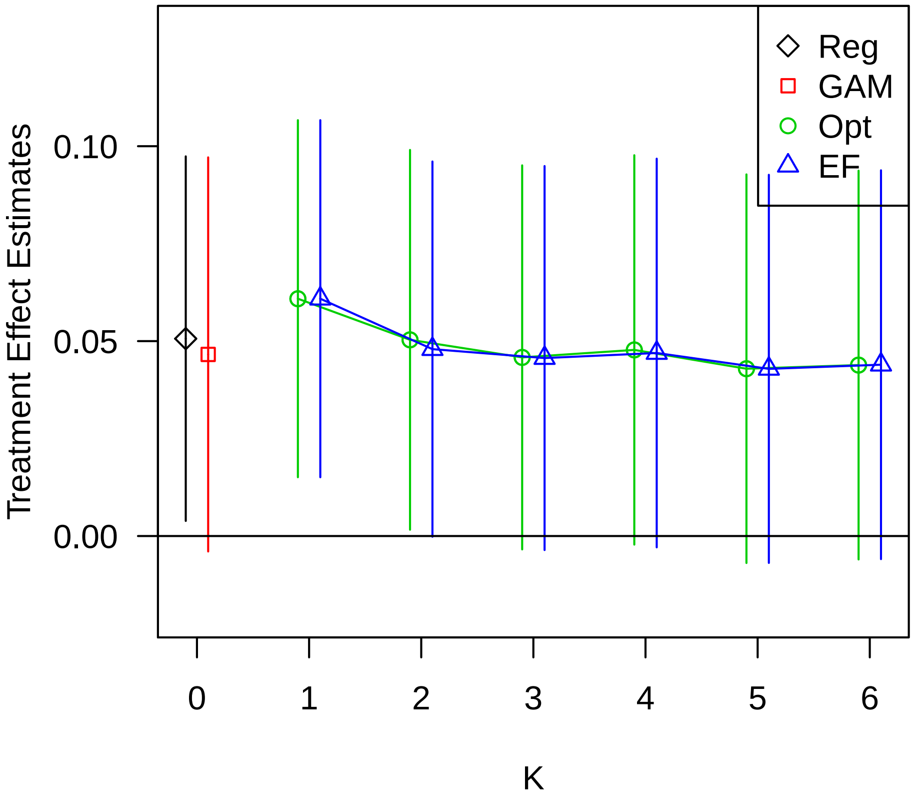

Figure 6. Treatment effect estimates with confidence intervals, using a regular regression approach (Reg), the GAM with propensity scores approach (GAM), and the optimal (Opt) and equal frequency (EF) stratification on propensity score approaches.