10-5-2007

OPTIMAL PROPENSITY SCORE STRATIFICATION

Jessica A. Myers
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, jamyers@jhsph.edu

Thomas A. Louis
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Optimal Propensity Score Stratification

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: Stratifying on propensity score in observational studies of treatment is a common technique used to control for bias in treatment assignment; however, there have been few studies of the relative efficiency of the various ways of forming those strata. The standard method is to use the quintiles of propensity score to create subclasses, but this choice is not based on any measure of performance either observed or theoretical. In this paper, we investigate the optimal subclassification of propensity scores for estimating treatment effect with respect to mean squared error of the estimate. We consider the optimal formation of subclasses within formation schemes that require either equal frequency of observations within each subclass or equal variance of the effect estimate within each subclass. Under these restrictions, choosing the partition is reduced to choosing the number of subclasses. We also consider an overalll optimal partition that produces an effect estimate with minimum MSE among all partitions considered. To create this stratification, the investigator must choose both the number of subclasses and their placement. Finally, we present a stratified propensity score analysis of data concerning insurance plan choice and its relation to satisfaction with asthma care.

KEY WORDS: Observational studies; Propensity score; Subclassification.

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 important confounding covariates, and this confounding biases the direct estimate of treatment effect (Rubin, 1991) (Sommer and Zeger, 1991). Methods available to control for confounding in observational studies include direct adjustment, matching, and stratification on covariates (Cochran, 1968) (Billewicz, 1965). Propensity score methods, developed by Rosenbaum and Rubin (1983), may also be used.

The propensity score of an experimental unit is the conditional probability of that unit being assigned to the treatment group, given observed covariates. Under randomization, this probability is controlled by the investigator and is independent of covariates, so that the propensity creates the assignments. When units are not randomized, the propensity emerges from the assignment process. Specifically, those who are treated will tend to have higher propensity scores than those who go untreated. This imbalance in propensity score represents an imbalance in covariates between treatment and control groups, and several methods utilizing the propensity, including matching, subclassification, and direct adjustment on propensity scores, have been shown to yield unbiased estimates of treatment effect (Rosenbaum and Rubin, 1984) (Rosenbaum and Rubin, 1983) (Rosenbaum and Rubin, 1985) (Dehejia and Wahba, 2002). We are interested in the method of adjustment by subclassification on propensity score.

When using the subclassification approach, the range of propensity scores is split into subclasses, and treatment effect is estimated for the outcomes within each subclass. The overall treatment effect is then estimated using an inverse variance weighted mean of the subclass-specific estimates. Choice of the number and placement of subclasses influences the variance and bias of the resultant estimate. Generally, there are opposing effects on variance and bias; wide classes produce low variance but high potential bias, narrow classes the reverse.

Two popular approaches for forming subclasses are equal frequency of observations in each subclass and equal subclass-specific variance of the estimated treatment effect across subclasses. In this paper we compare these two methods and a method designed to produce an approximately optimal set of strata, those that minimize the mean squared error (MSE) of the overall estimate. This hybrid approach depends on specifying a confounding structure, and near optimality of the approach depends on an approximately correct specification. In this report we base the partition on linear confounding. Mean squared error performance of the three approaches depends on the distribution of propensity scores in the two treatment groups, and we examine several patterns.

For a specified number of propensity score strata, the equal frequency approach often produces some subclasses with a large discrepancy between the number of units assigned to treatment and control, producing a high stratum-specific variance of the estimated treatment effect. Inverse variance weighting gives these estimates little influence on the overall estimate, in some sense "wasting" observations. The equal variance approach (Hullsiek and Louis, 2002) was designed to address this issue by forming strata with approximately equal variances for the estimated treatment effect. However, for a fixed number of subclasses, equal variance subclasses can have boundaries far from the equal frequency classes (generally, higher frequencies for low and high propensity scores, lower frequencies for propensity scores near 0.5).

Irrespective of the approach to forming class boundaries, MSE performance depends on the number of strata relative to the inherent degree of confounding. If there is no confounding, it is optimal to use one class (minimize the variance); if there is considerable confounding, then several strata will be needed to minimize MSE.

Section 2 describes the propensity score stratification methods, and Section 3 presents performance comparisons. Section 4 presents an analysis of an observational study of the effect of health insurance type on satisfaction with asthma care. Section 5 summarizes results and suggests another algorithm for choosing the number and form of subclasses.

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, attributes that help predict both treatment and outcome. 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. If the statistical relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is known, a standard covariate adjustment can be used to reduce bias. Alternatively, use of propensity score,

$$e(\boldsymbol{X_i}) = Pr(Z_i = 1 | \boldsymbol{X_i})$$

by weighting or stratification can reduce the bias (Rosenbaum and Rubin, 1984).

We study stratification into K subclasses ordered on propensity score, combining stratum-specific estimates. To structure the approach, let P_z be the propensity score random variable in treatment group z with density and cdf,

$$\begin{aligned} P_z &\sim f_z(p) \\ Pr(P_z \leq p) &= \int_0^p f_z(u)du \end{aligned}$$

For ease of notation we assume the distributions are continuous, however our findings hold for the general case. Determining an effective value for K and subclass boundaries depends on these distributions. For example, if $f_1(p) = f_0(p)$, then $K = 1$ is optimal (just use the overall estimate). Note that the $f_z(p)$ can be estimated and in some situations (e.g., a randomized experiment) are known.

Estimating the treatment effect with the simple difference of means permits the computation of subclass-specific variance, assuming without loss of generality that the variance of outcome in each group is 1. Let $\boldsymbol{t} = (0 = t_0 < t_1 < \ldots < t_K = 1)$ define a partition of the range of propensity scores with K subclasses. The estimated treatment effect in the k^{th} stratum and its variance are

$$\begin{aligned}
\hat{\Delta}_k = \hat{\Delta}_k(\boldsymbol{t}) &= \bar{Y}_{1k} - \bar{Y}_{0k} \\
V_k = V_k(\boldsymbol{t}) &= Var(\bar{Y}_{1k} - \bar{Y}_{0k}) \\
&= \frac{1}{N}\left[\frac{\pi^{-1}}{F_1(t_k) - F_1(t_{k-1})} + \frac{(1-\pi)^{-1}}{F_0(t_k) - F_0(t_{k-1})}\right].
\end{aligned}$$

where $N = n_0 + n_1$, and $\pi = n_1/N$. The overall estimate and it's variance are

$$\begin{aligned}
\hat{\Delta} = \hat{\Delta}(\boldsymbol{t}) &= \left(\Sigma_{k=1}^{K} \hat{\Delta}_k V_k^{-1}\right) / \left(\Sigma_{k=1}^{K} V_k^{-1}\right) \\
V(\hat{\Delta}) &= 1/\left(\Sigma_{k=1}^{K} V_k^{-1}\right).
\end{aligned}$$

If $V_k \equiv V$ (as in equal variance partitioning), then $V(\hat{\Delta}) = \frac{V}{K}$. Setting $K = 1$ yields the minimum variance, and variance increases with K. However, the bias is maximized at $K = 1$ and generally decreases as K increases.

In order to specify the bias, we must define the model,

$$Y|z, \boldsymbol{X} = \alpha + \beta z + \eta e(\boldsymbol{X}) + \epsilon \tag{1}$$

Here β is the true treatment effect, but we observe a linear combination of β and η. The within subclass bias and overall bias are given by

$$\begin{aligned}
E[\hat{\Delta}_k] - \beta &= (\alpha + \beta + \eta e_{1k}) - (\alpha + \eta e_{0k}) - \beta \\
&= \eta(e_{1k} - e_{0k}) \\
E[\widehat{\Delta}] - \beta &= \eta \frac{\Sigma_{k=1}^{K} V_k^{-1}(e_{1k} - e_{0k})}{\Sigma_{k=1}^{K} V_k^{-1}}
\end{aligned}$$

where $e_{zk} = E[P_z | P_z \in (t_{k-1}, t_k)]$.

After studying the formulas for variance and bias, it is clear that η is important in determining the mean squared error of the estimate associated with a particular subclass partition. For example, if either $\eta = 0$ or $e_{1k} - e_{0k} = 0$ for all k, then the estimate is unbiased, and one subclass is preferred. However, as η increases, the proportion of MSE attributable to bias increases, and so more subclasses are required in order to control MSE.

The number of units in the sample, N, also determines the relative impact of variance and bias in MSE, but because increasing sample size has the same effect on this relation as increasing η, we keep N constant and compare the various subclass formation methods at different levels of what we will call η^*, a representation of the combined effects of both η and N. Also note that in order to form K equal frequency or equal variance subclass partitions, it is not necessary to have any information about the value of η^*. However, in order to find the optimal subclass partition, which yields the minimum mean squared error estimate of treatment effect using K subclasses, η^* must be known.

3. Performance Assessment

To investigate the performance of the two methods for subclass formation already discussed compared to the optimal subclass formation, we created a program to compute the mean squared error at varying values of η using the variance and bias formulas above. The program considered all possible subclass partitions with cutpoints specified up to three decimal places, allowing between one and six total subclasses. It reported the lowest MSE of the six possible equal frequency formations, the lowest MSE of the six possible equal variance formations, and the number of subclasses used in each. The program also reported the subclass formation with the absolute minimum MSE of all partitions considered.

For simplicity, we require that the marginal density of propensity scores is uniform on $[0, 1]$, forcing equal frequency cutpoints to be equidistant along the range of propensity scores. We also assume that the densities of the propensity scores of the two treatment groups are antisymmetric, $f_0(p) = f_1(1-p)$, and that there is equal sample size in each treatment group. These restrictions yield

$$\begin{aligned} &\frac{1}{2}f_1(p) + \frac{1}{2}f_0(p) = 1 \qquad 0 < p < 1 \\ \Rightarrow \quad &f_1(1-p) = 2 - f_1(p) \qquad 0 < p < .5 \end{aligned} \tag{2}$$

which implies we must have $f_1(p) \leq 2$. Also notice that anti-symmetry causes the outermost regions of the propensity score range to have the greatest imbalance in group distribution, so if equal frequency subclass formation is used, the subclasses on each end of the range will have the largest variance.

Characterizing this class of functions, we may take any function f such that $f(p) \leq 2$ and $f(.5) = 1$ and define

$$f^*(p) = \begin{cases} f(p) & p \leq .5 \\ 2 - f(1-p) & p > .5 \end{cases}$$

Then, the function f^* satisfies the necessary conditions to be a density under consideration because $\frac{1}{2}f^*(p) + \frac{1}{2}f^*(1-p) = \frac{1}{2}[f(p) + 2 - f(p)] = 1$, as in 2. In fact, the form defined by f^* completely characterizes the class of functions that satisfy anti-symmetry and marginal uniformity. To show this, take any f in this class and define f^* as above. Then for $0 \leq p \leq .5$, $f^*(p) = f(p)$, and for $.5 < p \leq 1$, $f^*(p) = 2 - f(1-p) = f(p)$ since f satisfies (2).

For ease of computation, we begin with the functions defined by $f_1(p) = (2p)^s$, $f_0(p) = 2 - (2p)^s$. Varying s produces a wide range of plausible distributions, with an s of zero indicating a uniform distribution in each treatment group, and a large s indicating an extreme differential in treatment assignment according to propensity score.

3.1 *Results*

The program was run using the propensity score densities above with $s = 0.25$, $s = 1$, and $s = 3$. For visual aid, these densities are plotted below (Figure 1). The total sample size N was kept constant at 100 throughout, so that all change in η^* was captured by varying η ($\eta^* = \eta$). We plotted the mean squared errors of treatment effect estimates using the overall optimal, best equal frequency, and best equal variance subclass formations versus η^*, which ranged from zero to two. In addition, the number of subclasses used in each estimate was marked adjacent to the MSE of that estimate (Figure 2).

[Figure 1 about here.]

[Figure 2 about here.]

3.1.1 *Known η^*.* For smaller values of s, both equal frequency and equal variance subclass formations have nearly identical mean squared errors to the optimally formed subclasses. However, when $s = 3$ there is a clear benefit of optimal formation at intermediate values of η^*, with equal frequency formation having the second lowest values of MSE, and equal variance formation producing the highest values of MSE. This ordering was repeated at other values of s, with higher values of s generally producing larger differences in results between formation schemes.

As expected, the number of subclasses required to minimize the MSE of treatment effect estimation generally increased with increasing η^* under all three formation schemes and all three values of s, although not necessarily consecutively. In particular, among lower values of s ($s = 0.25$), an even number of subclasses is preferred in optimal subclass formation, and among higher values of s ($s = 3$), an odd number of subclasses is preferred. This finding is reasonable since at lower values of s the center of the conditional densities are nearly uniform, and therefore, stratifying this portion of the data is not beneficial in terms of variance or bias. In contrast, at higher values of s, the expected within interval difference in propensity scores is so great in the center of the distributions that not stratifying the center would severely increase bias. When $s = 1$, there is no apparent preference between even or odd numbers of subclasses.

Figure 2 also does not show consecutiveness of subclass count within even or odd counting, for example the jump from 1 to 5 among the overall optimal when $s = 3$ or the jump from 2 to 6 among $s = 0.25$. These jumps are due to the coarseness of the vector of values tried for η^*, and when η^* was further refined, intermediate numbers of subclasses were found where expected within the existing even or odd trend.

$3.1.2$ *Unknown η^*.* It should be noted that the optimal formations used above to estimate treatment effect assume that η^* is known. In practice, the η^* used to optimize the subclass formation will not necessarily be the true η^* under the model. Therefore, when comparing the performance of the various formation schemes, mean squared error should be calculated under the true η^*. We adjusted the program to assess the sensitivity of the three formation schemes to inaccurately estimated η^*. Subclass formations were created and optimized under our assumed η^*, but mean squared error of the resulting estimate was computed under one of four truths: $\eta^* = 0$, 0.5, 1, or 2.

[Figure 3 about here.]

Figure 3 presents the MSE of estimates under each of these four η^* values. From these plots it is clear that optimizing subclass formation for an inaccurate η^*, even within the equal frequency or equal variance schemes, can cause inflation of MSE. As in Figure 2, the effect is mitigated when the conditional densities of propensity scores are more uniform ($s = 0.25$), but may still result in an MSE as much as three times what would be expected when an appropriate number of subclasses are used. Also, it appears that underestimating η^* is more harmful than overestimating it, although the mean squared errors of estimates formed under an assumed η^* of near 2 may be conservative due to the fact that our program was limited to a maximum of six subclasses.

In summary, the overall optimal subclass formation has slightly lower MSE around the truth, as expected. However, at all other values of assumed η^*, none of the formation schemes are effective at reducing MSE because each chooses an incorrect number of subclasses when the estimated η^* is not close to the truth.

3.2 *Generality of Distributional Assumptions*

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

Let $F(p)$, $F_1(p)$, and $F_0(p)$ be the marginal and conditional cumulative distributions of propensity scores, respectively. Using the cdf as the uniform transform, the transformed scores and their conditional distributions are given by

$$\begin{aligned} F(p) &= \frac{1}{2}F_1(p) + \frac{1}{2}F_0(p) \\ &= U \sim U(0,1) \\ F_z^*(u) &= = F_z(F^{-1}(u)) \\ f_z^*(u) &= \frac{f_z(F^{-1}(u))}{f(F^{-1}(u))} \end{aligned}$$

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. Unfortunately, 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 partially generalizable to cases which do not meet the assumptions held thus far. We will now present one such case.

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 of confidentiality issues, and because we prefer a dichotomous 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 so we will consider the subset of data with these two insurance types so that the treatment of interest is dichotomous.

We began by developing a propensity score from the available covariates, including possible confounders such as age, sex, race, education, employment, and physician group. The propensity score was found through a logistic model of the personal health insurance indicator on the covariates of interest, so that the propensity score obtained is the probability of holding personally purchased insurance, rather than employer purchased insurance, controlling for all covariates in the model. After comparing several models through regular model checking and diagnostics, a model for generating propensity scores was selected to insure that the propensity score is accurate and that the model contains the apppropriate covariates, which at minimum includes the confounders, as suggested by Austin et al. (2007). In addition, we checked independence of treatment (health insurance type) and covariates, conditional on propensity score, 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. Finally, the propensity score was transformed to yield marginal uniformity as shown above.

[Figure 4 about here.]

[Figure 5 about here.]

Once a propensity score was established for each individual, we estimated η, the effect of propensity score on outcome, in order to begin subclass formation. We used simple covariate adjustment to produce a rough estimate through the model (1) where z now refers to the indicator of personally purchased insurance. Although this is a reasonable way to estimate η, it is not sufficient here to accurately estimate the treatment effect without the possibility of considerable bias because it relies heavily on model assumptions of a linear and additive relation between treatment and response.

It should be noted that this procedure is only valid when the relation between propensity score and response is monotone. Otherwise, a U-shaped relationship could result in an η estimate of 0, indicating no bias, despite a large bias term. The value of η estimated with the above model is 0.124. This estimate, combined with the imbalance in propensity between treatment groups observed in the plot of conditional cumulative distributions (Figure 5), indicates that there may be a sizeable bias term on the naive estimate of treatment effect.

[Figure 6 about here.]

[Figure 7 about here.]

In order to examine the performance of the equal frequency subclass formation scheme as the subclass count is increased, we plotted the within subclass estimates and the overall estimate of treatment effect, using each of one through ten subclasses (Figure 6). The within subclass estimates were computed using a simple difference of means (unadjusted), as well as the linear regression of outcome on treatment and propensity (adjusted). Also reported is the p-value from a χ^2 test of equality among within subclass linear regression estimates.

Comparing the difference of means estimate and linear regression estimate using $K = 1$ (no stratification), bias due to treatment assignment is apparent. Increasing the number of subclasses to two reduces the difference between adjusted and unadjusted estimates, and using three or more subclasses causes the overall difference of means and linear regression estimates to be nearly equal, indicating a reduction in bias. The two overall estimates remain nearly unchanged as the number of subclasses is increased from three, and the increase in variance on these estimates is negligible, leading to essentially identical inferences when any of three through ten subclasses are used (Figure 7).

The adjusted and unadjusted estimates within a subclass are also nearly identical beginning with three subclasses, indicating that the variation of propensity scores within a subclass is not substantially effecting the estimate. Therefore, subclassification is effectively reducing treatment assignment bias. However, there is an apparent trend across subclasses which produces generally higher effect estimates in subclasses with higher propensity scores. The trend in estimates is mitigated by $K = 9$, so if the overall estimates were changing as K increased, one might prefer the stratification with nine or ten subclasses.

Figure 6 shows treatment effect estimates in subclasses with lower propensity scores have larger variances. This trend is different from the anticipated obstacle of very large variances in the outer subclasses, but it does not represent a major concern, since the differential among variances is not great at moderate values of K $(K < 6)$. In general, the estimated variance of within subclass estimates, as well as their observed variability from one subclass to the next, increases as K increases, but there is no trend in χ^2 p-values as the number of subclasses increases.

After examining all of the equal frequency analyses presented above, we recommend using the linear regression estimate generated from the three subclass partition. We choose the linear regression estimate because it is much more stable at varying K than the difference of means estimate. We choose the three subclass partition because it contains the minimum number of subclasses necessary to suitably control bias in treatment effect estimates, and it does not contain enough subclasses to increase variance. Also, three subclasses are chosen because the relative uniformity of the conditional densities of propensity scores (Figure 4) indicate a preference for an odd number of subclasses, as reported in the performance assessment.

Athough this method was not investigated in the performance assessment, we also considered using the within subclass estimates of bias to guide fractionation of subclasses. This adaptive fractionation method would further split, according to frequency, subclasses with bias estimates above some threshold, and leave remaining subclasses unsplit. Bias may be estimated using the rough estimate of η reported above and the difference in mean propensity scores between treatment and control groups.

Choosing a threshold of $.005$, the adaptive fractionation approach indicates a need for more than one subclass in this example, since the estimated bias for $K = 1$ is $.032$. Therefore, we consider $K = 2$. Both of the within subclass bias estimates in this partition are above our threshold, so rather than splitting either subclass in half, consider $K = 3$. All subclasses in the three subclass partition yield bias estimates below our designated threshold, so we stop at $K = 3$. If one or two of the bias estimates from this partition had been above $.005$, we would have split those subclasses in half. If all three bias estimates had been above $.005$, we would have considered $K = 4$. It is fortunate that in this analysis the adaptive fractionation approach leads to the same stratification already decided upon: three equal frequency subclasses.

We must note that, as with the model used above to estimate η, linear regression was used throughout this analysis, despite the dichotomous outcome measurements of satisfaction with care (poor/fair/good vs. very good/excellent) because the results thus far relied on a linear bias structure. Since the superiority of equal frequency formation over equal variance formation was established in the performance assessment in simple linear cases such as this example, we did not pursue equal variance stratification. Also, the conditional densities of propensity resemble those considered in the performance assessment with $s < 1$, indicating that an appropriate number of equal frequency subclasses would control for bias approximately as well as the optimal subclass formation, so we did not pursue optimal stratification in this example either.

5. Discussion and Future Work

The objective of this study was to compare the effectiveness of equal frequency and equal variance partitioning schemes at controlling the mean squared error of the effect estimate in a stratified propensity score analysis, and to consider the relative merits of an optimal partition. Although there was no uniformly best strategy, the results imply some practical suggestions in choosing propensity score subclasses.

First, when outcome follows a linear additive model, as was considered here, equal frequency subclass formation results in effect estimates with consistently lower MSE than equal variance formation, regardless of accuracy in the estimation of η. Also, the use of an appropriate number of equal frequency subclasses was in general not substantially inferior to the optimal subclass partition, and sometimes resulted in lower MSE of the estimate compared to the optimal when subclasses were chosen under an inaccurately estimated η^*. Therefore, considering the additional computational expense of the equal variance and overall optimal methods, equal frequency formation is recommended in this case. However, when using within subclass regression adjusting for variables other than propensity score to estimate treatment effects, the equal variance method may still perform better than equal frequency with respect to MSE, as shown by Hullsiek and Louis (2002).

Second, the appropriate choice of partition relies on the relative importance of bias and variance in the MSE, summarized by η^*, and the level of imbalance in propensity to assignment between the two treatment groups, summarized in this study by s. As η^* increases, more subclasses are needed to effectively control the increasing impact of the bias term, and using more subclasses than is optimal in equal frequency and equal variance formations is less harmful to MSE than using too few subclasses. Thus, when using the equal frequency method to form subclasses, we suggest increasing the number of subclasses until the overall estimate remains relatively constant, or until there is a substantial increase in variance of the estimate. Also, the optimal subclass formations, both within a scheme and overall, generally stratified sections of the domain of propensity scores with large slopes on the conditional densities in those sections. This behavior is due to the large values of $e_{1k} - e_{0k}$ found in subclasses that span these sections, and it explains the preference for odd K when the edges of the conditional densities have high slope ($s < 1$), and the preference for even K when the centers of the densities have high slope ($s > 1$). One should consider this preference when choosing subclasses to ensure that these sections of propensity score domain are appropriately stratified in order to control bias.

A possibility for achieving overall optimal partitioning is the adaptive fractionation pursued in the example analysis. The estimated relation between propensity score and outcome through direct covariate adjustment is a competent guide for the choice of an initial K equal frequency subclasses, and comparing estimated bias across the K subclasses indicates which subclasses may need further fractionation. Adaptively splitting subclasses with this algorithm will provide better variance control than the purely equal frequency method, but it creates additional dependence on correct model choice and could dictate poor splitting choices under an incorrect model. Combining within subclass covariate adjustment with adaptive fractionation would allow for maximum flexibility in the true underlying model and robustness from either strategy.

Finally, both the performance assessment and the example analysis indicate that equal frequency subclass formation adequately controls MSE of the estimate of treatment effect, and an improved method for generating subclasses is necessary only in cases of extreme imbalance in propensities between treatment groups. Our results are without loss of generality with respect to the assumption of uniformity on the propensity score distribution, but that generality may not extend to conditional propensity score distributions which do not satisfy anti-symmetry, since no monotone transformation exists to generate anti-symmetry. Furthermore, these methods must be reevaluated for application to nonlinear data.

Acknowledgements

This research was supported by Grant 5T32ES012871 from the U.S. National Institute of Environmental Health Sciences. The authors wish to thank 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.

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

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

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

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

Appendix Proof of Anti-symmetry Preservation

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

This paper has been typeset from a TEX/ LATEX file prepared by the author.

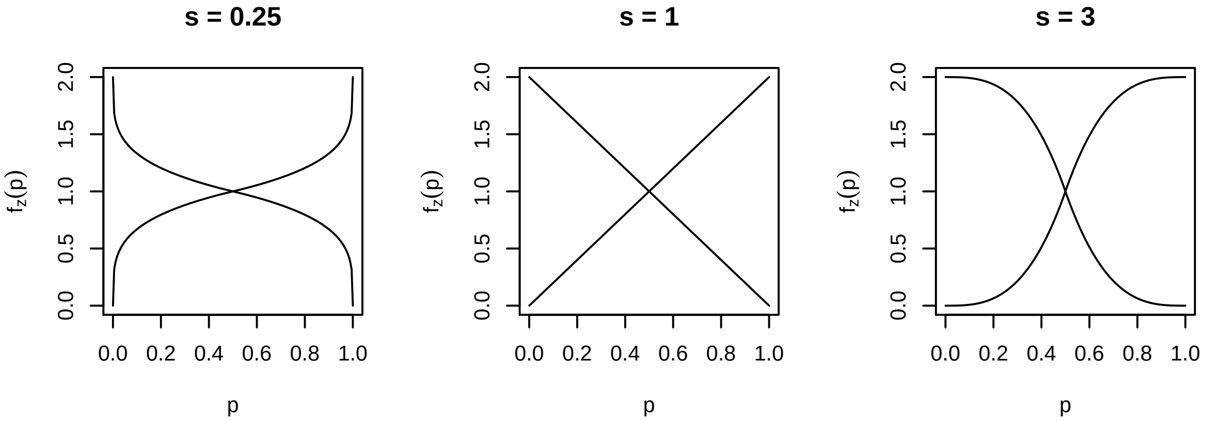

Figure 1. The three pairs of propensity score densities conditional on treatment that were considered.

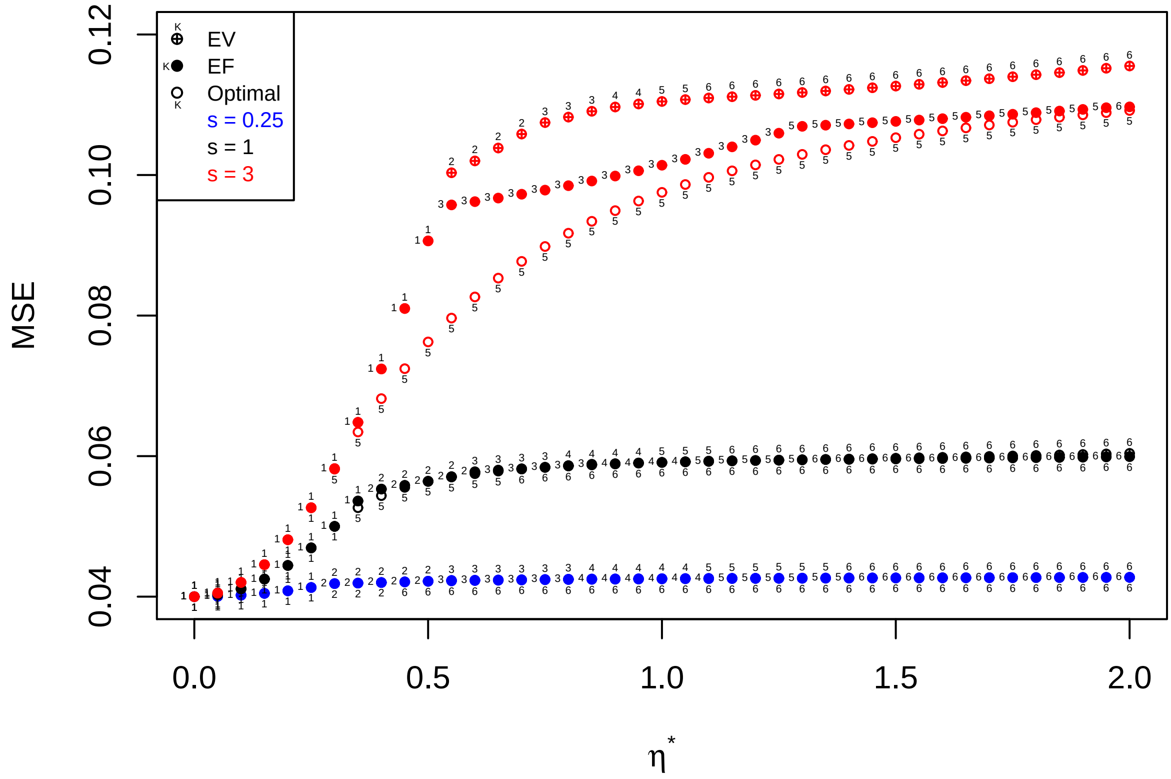

Figure 2. Mean squared error of optimal estimates within each subclass formation category: equal frequency, equal variance, and overall optimal, with the number of propensity score subclasses used in that estimate adjacent. These are plotted for each conditional density pair shown above: s = 0.25, s = 1, s=3.

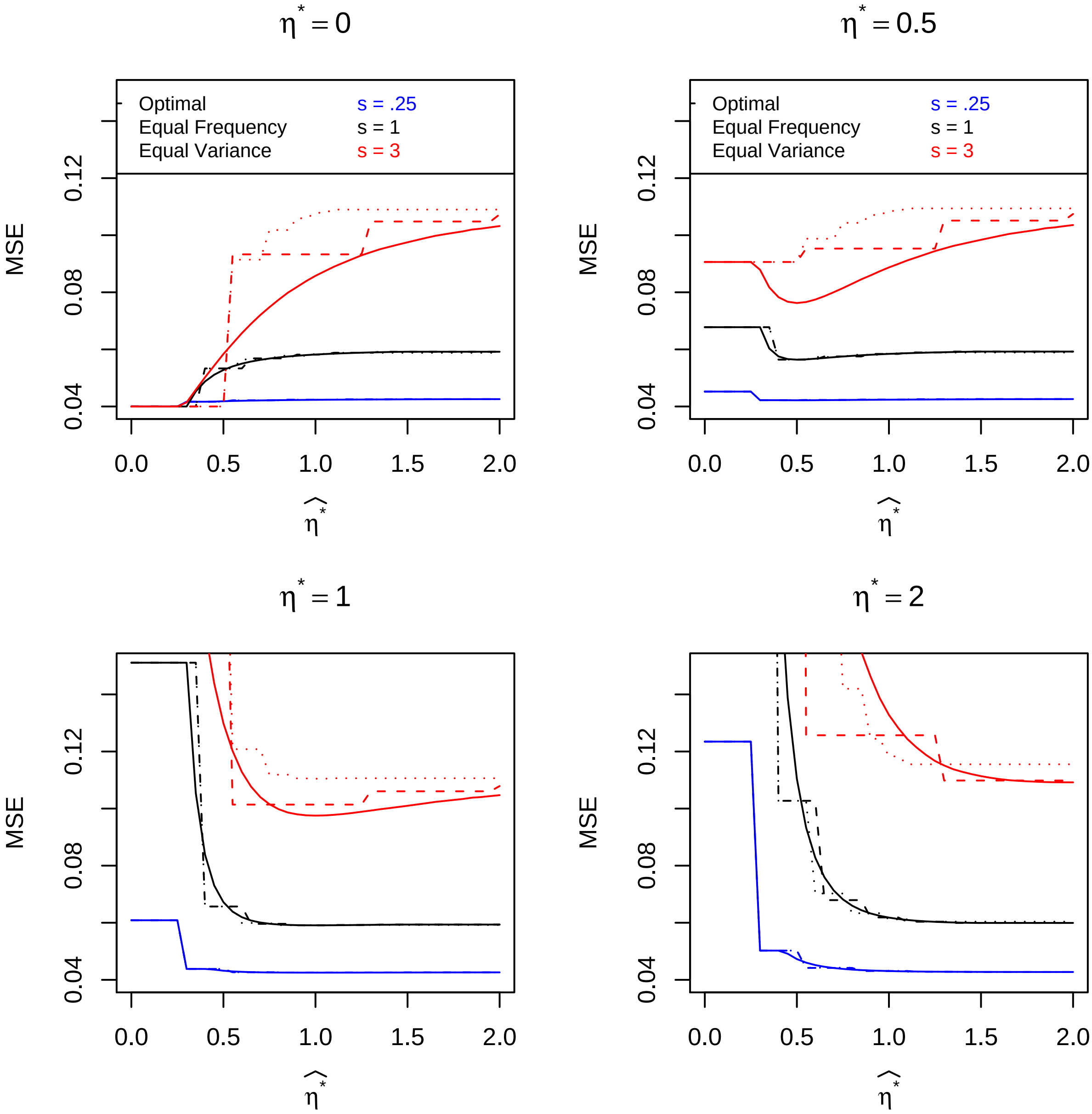

Figure 3. MSE of treatment effect estimators at each value of true η^* considered, using propensity score subclasses formed optimally within one of three formation schemes under the assumed $\widehat{\eta^*}$ and three propensity score conditional densities.

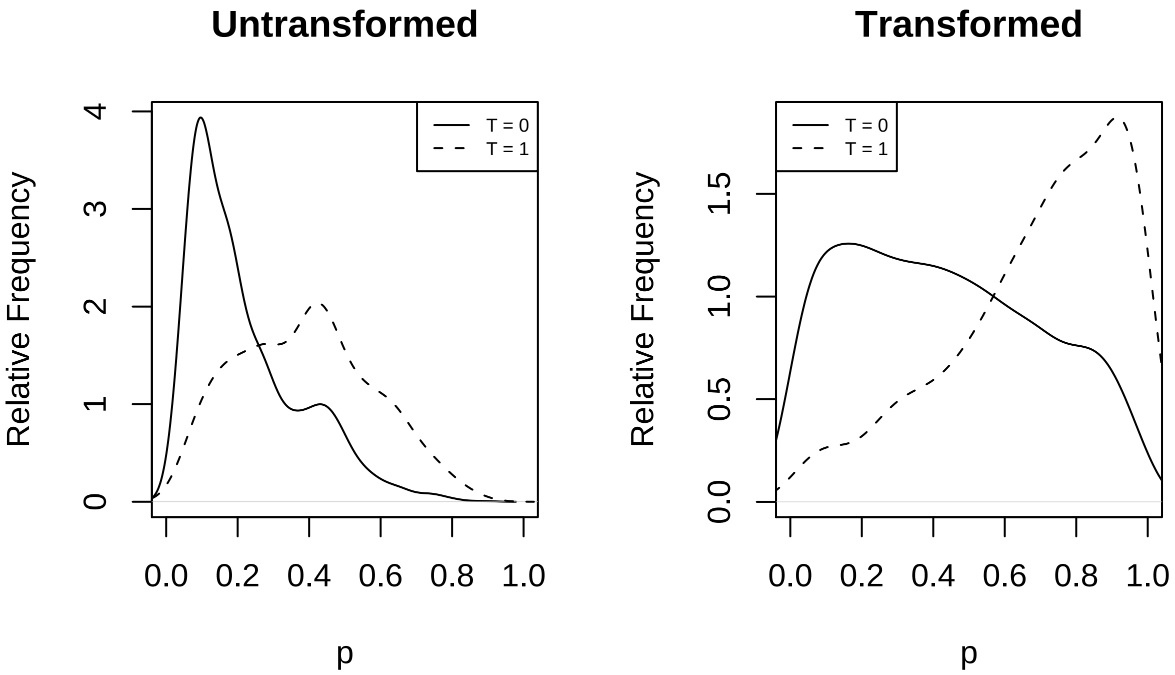

Figure 4. Relative frequencies of the raw and uniform transformed propensity scores conditional on treatment. Only 26.4% of units had personally purchased health insurance (T=1), and 73.6% of units had employer purchased health insurance (T=0).

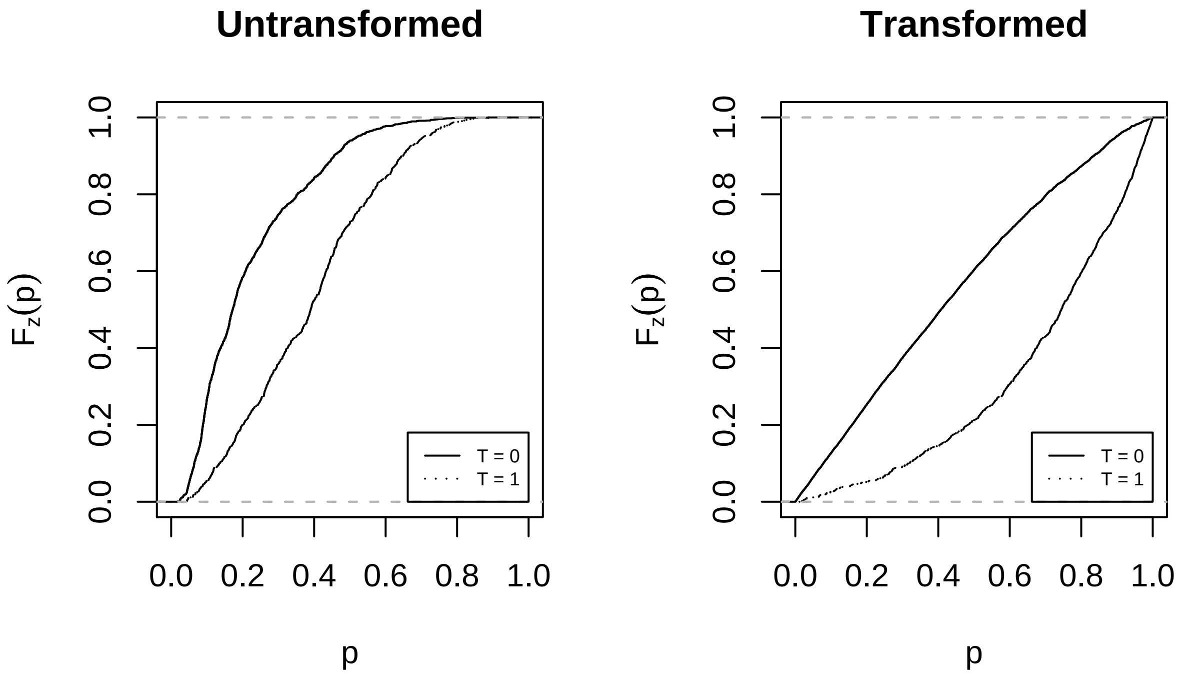

Figure 5. Cumulative distributions of raw and uniform transformed propensity scores conditional on treatment. Only 26.4% of units had personally purchased health insurance (T=1), and 73.6% of units had employer purchased health insurance (T=0).

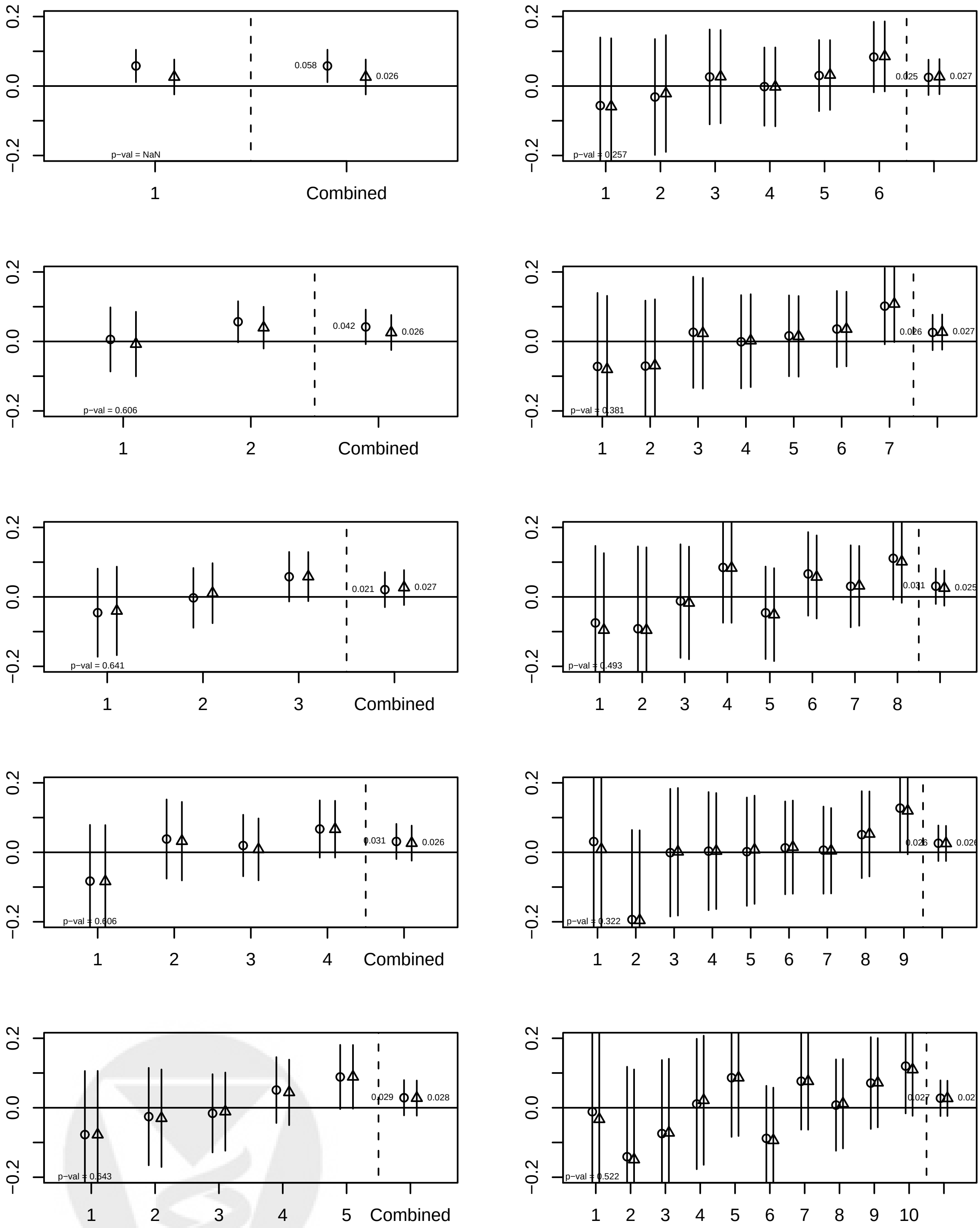

Figure 6. Within subclass and overall treatment effect estimates with confidence intervals, using an equal frequency subclass formation of between one and ten subclasses. A circle represents an estimate derived from the difference of means and a triangle represents an estimate derived from the linear regression model. The p-value from a χ^2 test for independence of the K within subclass linear model estimates is reported in the lower right corner.

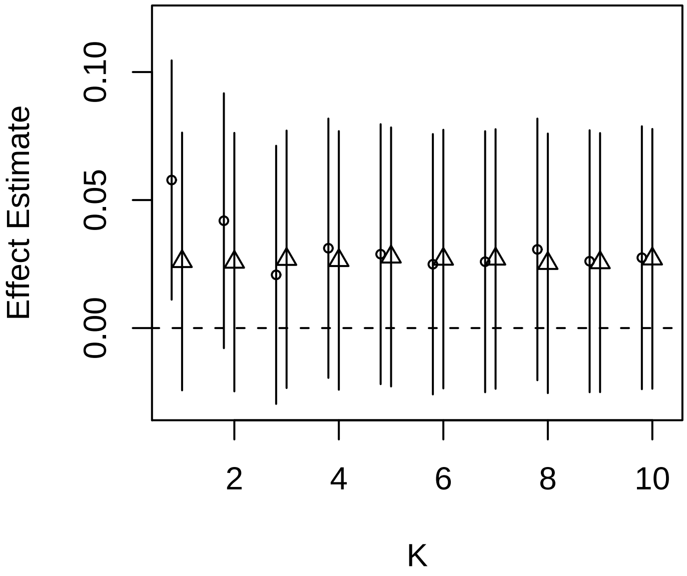

Figure 7. Overall treatment effect estimates with confidence intervals, using an equal frequency subclass formation of between one and ten subclasses. A circle represents an estimate derived from the difference of means and a triangle represents an estimate derived from the linear regression model.