Chapter 17
Subclassification on the Propensity Score

17.1 Introduction

In this chapter we discuss a method for estimating average treatment effects under unconfoundedness based on subclassification on the propensity score. We refer to this method also as blocking or stratification, Recall the definition of the propensity score in Chapter 12 as the conditional probability of receiving the treatment given the covariates. The propensity score is a member of a class of functions of the covariates, collectively referred to as balancing scores, that share an important property. Within subpopulations with the same value for a balancing score, the distribution of the covariates is identical in the treated and control subpopulations. This in turn was shown to imply that, under the assumption of unconfoundedness, biases associated with observed covariates can be eliminated entirely by adjusting solely for differences between treated and control units in a balancing score. The practical relevance of this result stems from the fact that a balancing score may be of lower dimension than the original covariates. (By definition the covariates themselves form a balancing score.) When the balancing score is of lower dimension than the full set of covariates, effectively adjusting for differences in the balancing score may be easier to implement than adjusting for differences in all covariates, by avoiding high-dimensional non-parametric methods. Within the class of balancing scores the propensity score, as well as strictly monotonic transformations thereof (such as the log odds ratio), have a special place. Since the propensity score is a scalar function of the covariates, there can be no balancing score of lower dimension. It thus maximizes the dimension reduction relative to using the original covariates.
In this chapter we will examine a leading approach to exploiting the balancing property of the propensity score for the purpose for estimation of average treatment effects. The method relies on blocking on the propensity score. The sample is divided in subclasses (strata, or blocks), based on the value of the propensity score, so that within the subclasses the propensity score is, at least approximately, constant. We then estimate the average treatment effect within each subclasses as if assignment was completely random, using either the Neyman-based methods for completely randomized experiments from Chapter 6, or the regression and model-based methods from Chapters 7 and 8. To estimate the overall average treatment effect we average the within-block estimated average treatment effects, weighted by the subclass sizes. We can estimate other estimands using the model-based methods from Chapter 8. Two important practical issues in implementing this method are, first, the choice of the number of subclasses or blocks, and, second, the choice of boundary values for the blocks.

Often we combine subclassification with further covariance adjustments, and in fact we generally recommend doing so. Such adjustments have two objectives. First, because blocking typically does not eliminate the entire bias associated with differences in the covariates, regression or model-based adjustments can further reduce the bias. Second, these adjustments can improve the precision of estimators for treatment effects. There is an important difference though, between the covariance adjustment in this setting, where it is combined with propensity score methods, and its use in the full sample. In the latter case there is generally concern that the implicit imputations of the missing potential outcomes through regression methods rely heavily on extrapolation. Here, by the construction of the strata, the difference in covariate distributions will be smaller, often much smaller, and extrapolation is less of an issue.

In the next section we return to the Imbens-Rubin-Sacerdote lottery data, previously used in Chapter 14, that will be used to illustrate the concepts discussed in this chapter. Then we return to theoretical issues. In Section 17.3 we discuss the construction of subclasses, and the bias reduction properties of these methods. In the following section, Section 17.4, we implement this on the lottery data. In Sections 17.5 and 17.6 we develop estimators for average treatment effects based on subclassification and methods for inference. These methods are then implemented on the lottery data in Section 17.7. In Section 17.8 we discuss the relation to weighting methods. We conclude in Section 17.9.
17.2 The Imbens, Rubin and Sacerdote Lottery Data

In this chapter we use the same lottery data originally collected by Imbens, Rubin and Sacerdote (2001) that we used as one of the illustrations in Chapter 14. In Chapter 14 we assessed the overlap in covariate distributions. For the lottery data we found that overlap was substantial, although there were subsets of covariate values with little overlap. The first column in Table 1 presents the normalized differences for the full sample. Note that the normalized difference for the covariate \# Tickets is 0.64, suggesting simple linear regression may not be adequate to reliably remove biases associated with differences in this covariate. To address these concerns with overlap in covariate distributions we apply the methods discussed in Chapter 16 designed to improve the overlap by dropping units with values of the estimated propensity score close to zero or one. Following the specific recommendations from that chapter suggests dropping units with estimated propensity score values outside the interval \([0.0891, 0.9009]\). Table 2 presents the subsample sizes in the various propensity score strata. Out of the 496 units in the full sample, 259 losers and 237 winners, there are 323 with estimated propensity scores in the interval \([0.0891, 0.9009]\), of whom 172 are losers and 151 are winners. There are 86 observations discarded because of low propensity score values (less than 0.0891), 82 losers and 4 winners, and 87 discarded because of high propensity score values (higher than 0.9009), 5 losers and 82 winners.

On this trimmed sample we re-estimate the propensity score using the algorithm discussed in Chapter 13, for selecting linear and second order terms. Starting with the four variables selected for automatic inclusion, \# Tickets, Education, Working Then, and Earnings Year \(-1\), the algorithm selects four additional linear terms, Age, Pos Earnings Year \(-5\), Year Won, and Earnings Year \(-5\). In addition the algorithm selects four second order terms, Year Won \(\times\) Year Won, Tickets Bought \(\times\) Year Won, Tickets Bought \(\times\) Tickets Bought, and Working Then \(\times\) Year Won. Table 3 presents the parameter estimates for the logistic specification choosen. This is the estimated propensity score we work this for the purpose of this chapter. Note that when we used the same algorithm on the full sample we ended up with more terms, to be precise eight linear terms and ten second order terms. The trimming has improved the balance, and as a result the algorithm for specifying the propensity score selects fewer terms to adequately fit the data.
17.3 Subclassification on the Propensity Score and Bias Reduction

In Chapter 12 we showed that if unconfoundedness holds, it was sufficient to adjust for differences in the propensity score, or in fact for differences in any balancing score, to eliminate biases in comparisons between treated and control units. In principle we could build a statistical model for the potential outcomes given the propensity score, similar to the models discussed in the chapters on regression and model-based imputation in randomized experiments, Chapters 7 and 8. Since we do not have the complications arising from the presence of multiple covariates, modeling might appear even more straightforward than it was in those chapters. However, there are some questions concerning modeling the conditional distribution, or the expectation of, the potential outcomes given the propensity score (or any other balancing score), that make standard modeling approaches involving smooth flexible functional forms less attractive. The balancing scores are functions of the covariates with particular statistical properties, but without a substantive interpretation. It is therefore not clear what statistical models are appropriate for the conditional distribution of the potential outcomes given a balancing score. In particular there is no reason to suppose that a Gaussian linear model, even after some transformation of the outcome or the balancing score, will give a good approximation to the conditional distribution.

For this reason we do not start by building models in the same way we did in Chapters 7 and 8, using linear models and extensions thereof. Rather, we attempt to build relatively flexible models that do not require much substantive knowledge of the conditional distribution of the potential outcomes given the balancing score. In this chapter we do so by classifying or stratifying units by a coarsened value of the propensity score, similar to the way we used propensity score strata in Chapter 13 to evaluate the specification of the propensity score. Note that the construction of strata based directly on the full set of covariates would be infeasible with large number of covariates, as the number of subclasses that would be required to make the variation in covariates within subclasses modest, would generally be very large. For example, with the eighteen covariates in the lottery example, even if we defined subclasses in terms of just two (ranges of) values of the covariates would lead to an infeasibly larger number of subclasses, namely $2^{18} = 262144$, much larger than the sample size.
17.3.1 Subclassification

As in Chapter 13, let us partition the range of the propensity score into $J$ blocks, that is, intervals of the type $(b_{j-1}, b_j]$, where $b_0 = 0$ and $b_J = 1$ so that $\cup_{j=1}^J (b_{j-1}, b_j] = (0, 1]$. We shall analyze the data within a stratum as if they arose from a stratified randomized experiment. This means that we shall analyze units with propensity scores within an interval $(b_{j-1}, b_j]$ as if they have identical propensity scores. For large $J$, and choices for the boundary values of the intervals so that $\max_{j=1,\ldots,J} |b_j - b_{j-1}|$ is at least moderately small, this may be a reasonable approximation.

Recall the notation from Chapter 13: for $i = 1, \ldots, N$, and for $j = 1, \ldots, J-1$, the binary stratum indicators $B_{ij}$ are defined as

$$B_{ij} = \begin{cases} 1 & \text{if } b_{j-1} \leq \hat{e}(X_i) < b_j, \\ 0 & \text{otherwise}, \end{cases}$$

and $B_{iJ} = \begin{cases} 1 & \text{if } b_{J-1} \leq \hat{e}(X_i) \leq b_J \\ 0 & \text{otherwise}. \end{cases}$

Also, the number of units of each treatment type in each strata is denoted by $N_{wj} = \sum_{i: W_i = w} B_{ij}$, for $w = c, t$, and $j = 1, \ldots, J$, and $N_j = N_{cj} + N_{tj}$.

Let $p_j$ be the proportion of treated units in stratum $j$, and let $q_j$ be the fraction of units in stratum $j$:

$$p_j = \frac{N_{tj}}{N_j}, \quad \text{and} \quad q_j = \frac{N_j}{N}, \quad \text{for } j = 1, \ldots, J.$$

We implement the selection of boundary points in the same iterative procedure as in Chapter 13, with slightly different cutoff values. We start with a single block: $J = 1$, with boundaries equal $b_0 = 0$ and $b_J = b_1 = 1$. We then iterate through the following two steps. In the first step we assess the adequacy of the current number of blocks. This involves calculating, for each stratum, a t-statistic for the null hypothesis that the average value of the estimated propensity score, or rather the log odds ratio, is the same for treated and control units in that stratum. The specific t-statistic used is

$$t_{\ell,j} = \frac{\overline{\ell}_{1j} - \overline{\ell}_{0j}}{\sqrt{S^2_{\ell,j} \cdot (1/N_{0j} + 1/N_{1j})}},$$

where $\overline{\ell}_{wj} = \frac{1}{N_{wj}} \sum_{i: W_i = w} B_{ij} \cdot \hat{e}(X_i)$, and

$$S^2_{\ell,j} = \frac{1}{N_{1j} + N_{0j} - 2} \left( \sum_{i: W_i = 0} B_{ij} \cdot (\hat{e}(X_i) - \overline{\ell}_{0j})^2 + \sum_{i: W_i = 1} B_{ij} \cdot (\hat{e}(X_i) - \overline{\ell}_{1j})^2 \right).$$
In addition we check for each stratum the number of treated and control units left in each sub-stratum after a subsequent split, where we would split the current stratum at the median value of the estimated propensity score. Specifically we check whether the number of controls and treated, \( N_{1j} \) and \( N_{0j} \), and the total number of units, \( N_j \), in each new stratum, are above some minimum cutoff value. If at least one of the stratum is not adequately balanced, and if splitting that stratum would lead to two new strata with a sufficient number of units, that stratum is split and the new strata are assessed for adequacy. In order to implement this algorithm we need to specify three parameters, the maximum acceptable t-statistic (\( t_{\text{max}} \)), and the minimum number of treated or control units in a stratum, \( N_{\text{min},1} \), and the minimum number of units in a new stratum, \( N_j \geq N_{\text{min},2} \). Here we choose \( t_{\text{max}} = 1.96 \), and the number of \( N_{\text{min},1} = 3 \), and \( N_{\text{min},2} = K + 2 \), where \( K \) is the dimension of the vector of covariates \( X_i \). Compared to the values we choose when we were assessing the adequacy of the propensity score itself, \( t_{\text{max}} = 1 \), and \( N_{\text{min},1} = 3 \), and \( N_{\text{min},2} = K + 2 \), we allow the blocks here to be less balanced, since we typically will use additional covariate adjustment procedures.

### 17.3.2 Subclassification and Bias Reduction

To gain insights into the properties and benefits of subclassification, we investigate here some implications for bias. In this discussion we build on the theoretical analysis of the bias reducing properties of matching presented in Chapter 15. As in the discussion in Chapter 15, we initially assume that the conditional expectation of the two potential outcomes given the covariates is linear in the covariates, with identical slope coefficients in both regimes:

\[
\mathbb{E}[Y_i(w) | X_i = x] = \alpha_w + \beta' x = \alpha_0 + \tau \cdot 1_{w=1} + \beta' x,
\]

for \( w = 0,1 \), where \( \tau = \alpha_1 - \alpha_0 \). As before, we do not believe this assumption is a good approximation (in fact, if it were true, one could simply remove all biases associated with the covariates by covariance adjustment), but linearity provides a useful approximation to assess the bias reducing properties of subclassification.

Now consider the full sample. Let \( \bar{X}_0 \), \( \bar{X}_1 \), and \( \bar{X} \), be the average values of the covariates in the control, treated and full samples respectively,

\[
\bar{X}_0 = \frac{1}{N_c} \sum_{i: W_i = 0} X_i, \quad \bar{X}_1 = \frac{1}{N_t} \sum_{i: W_i = 1} X_i \quad \text{and} \quad \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i = \frac{N_c}{N} \cdot \bar{X}_0 + \frac{N_t}{N} \cdot \bar{X}_1.
\]
In addition, let \( \overline{X}_{0j}, \overline{X}_{1j}, \) and \( \overline{X}_j \) denote the covariate averages within stratum \( j \),

\[
\begin{align*}
\overline{X}_{0j} &= \frac{1}{N_{0j}} \sum_{i:W_i=0} B_{ij} \cdot X_i, \\
\overline{X}_{1j} &= \frac{1}{N_{1j}} \sum_{i:W_i=1} B_{ij} \cdot X_i, \\
\overline{X}_j &= \frac{1}{N_j} \sum_{i=1}^N B_{ij} \cdot X_i,
\end{align*}
\]

for \( j = 1, \ldots, N \). First we consider the estimator with no adjustment for differences in the covariates at all, where simply estimate the average treatment on the full sample, without subclassification, by differencing the average outcomes for treated and control units. Alternatively this can be viewed as a subclassification estimator with a single stratum. We find

\[
\hat{\tau}_{\text{no-adj}} = \overline{Y}_1 - \overline{Y}_0 = \frac{1}{N_t} \sum_{i:W_i=1} Y_i - \frac{1}{N_c} \sum_{i:W_i=0} Y_i.
\]

The bias of this estimator, conditional on the covariates, has two components. First, we estimate the average outcome given the treatment for the \( N_c \) control units, in expectation equal to \( E[Y_i(1)|W_i = 0] \) by \( \overline{Y}_1 \), by is similar. It consists of the average outcome for \( N_t \) treated units, which is in expectation equal to \( E[Y_i(1)|W_i = 1] \). The second component consists of the difference in expected outcomes for the control outcomes for the treated units, \( E[Y_i(0)|W_i = 1] \) and the expected value of its estimator, the average of the control outcomes for the control units, \( E[Y_i(0)|W_i = 0] \). Hence the conditional bias of \( \hat{\tau}_{\text{no-adj}} \) is, under the linear model specification for the regression function, equal to:

\[
E[\hat{\tau}_{\text{no-adj}}|X, W] - \tau = \frac{N_0}{N} \cdot (E[Y_i(1)|W_i = 1] - E[Y_i(1)|W_i = 0])
-
\frac{N_1}{N} \cdot (E[Y_i(0)|W_i = 1] - E[Y_i(0)|W_i = 0])
=
\frac{N_c}{N} \cdot \beta'(\overline{X}_1 - \overline{X}_0) - \frac{N_t}{N} \cdot \beta'(\overline{X}_c - \overline{X}_t)
=
((1 - p) \cdot \beta + p \cdot \beta') \cdot (\overline{X}_t - \overline{X}_c) = \beta' \cdot (\overline{X}_t - \overline{X}_c),
\]

where, as defined before, \( p = N_t/N \), the fraction of treated units in the full sample.

Now consider estimating the average treatment \( \tau \) by subclassification, still with no further covariance adjustment within the strata. Suppose we use \( J \) subclasses, instead of one, with \( N_j = q_j \cdot N \) units in stratum \( j \). In stratum \( j \) the bias is, using the same argument as for the overall bias,

\[
E[\hat{\tau}_{j,\text{no-adj}}|X, W] - \tau_j = \beta'(\overline{X}_{1j} - \overline{X}_{0j}).
\]
The overall bias for the subclassification estimator is the weighted average of the within-block biases,

\[ \mathbb{E} [\hat{\tau}_{\text{block, no-adj}} | \mathbf{X}, \mathbf{W}] - \tau = \beta' \left( \sum_{j=1}^{J} q_j \cdot (\overline{X}_{1j} - \overline{X}_{0j}) \right). \]

As a result of the subclassification, the bias that can be attributed to differences in \( X_{i,k} \), the \( k \)-th element of the covariate vector \( X_i \), is reduced from

\[ \beta_k \cdot (\overline{X}_{1j,k} - \overline{X}_{0j,k}) \]

to

\[ \beta_k \cdot \left( \sum_{j=1}^{J} q_j \cdot (\overline{X}_{1j,k} - \overline{X}_{0j,k}) \right), \]

where \( \overline{X}_{wj,k} \) is the \( k \)-th element of \( \overline{X}_{wj} \). Thus, the bias attributable to covariate \( X_{ik} \) is reduced by a factor

\[ \gamma_k = \frac{\sum_{j=1}^{J} q_j \cdot (\overline{X}_{0j,k} - \overline{X}_{1j,k})}{(\overline{X}_{1, k} - \overline{X}_{0,k})}. \]  (17.1)

We can calculate these ratios \( \gamma_k \) for any particular subclassification, for each covariate, to assess the bias reduction from the subclassification in a particular application.

### 17.4 Subclassification and the Lottery Data

Here we return to the lottery data and determine the number of subclasses or strata according to the algorithm discussed in Section 17.3. We use the cutoff values \( \tau_{\text{max}} = 1.96 \), and \( N_{\text{min},1} = 3 \), and \( N_{\text{min},2} = K+2 \), where \( K \), the number of covariates, is 18, so that \( N_{\text{min},2} = 20 \). These choices for the tuning parameters lead to five blocks. The details for the five blocks, including the cutoff values for the propensity score, the number of units by treatment status in each block, and the t-statistics for the null hypothesis of a zero difference in average propensity scores between treated and control units in the block, are presented in Table 4. For example, the first stratum contains 67 control and 13 treated units, with the propensity scores ranging from 0.03 to 0.24. The t-statistic for the null hypothesis of no difference is -0.1, so there is actually very little difference in average propensity scores between the two groups within the first block.

For comparison purposes we also present, in Table 5, results based on two blocks, where the block boundary is the median value of the propensity score, 0.44. Here the groups are substantially less balanced.
Next, we investigate for these two specifications of the blocks, the extent of the bias reduction based on a linear specification of the regression function. In Table 1, we presented in columns 1 and 3, for both the full and trimmed samples, the difference in covariates, $\bar{X}_{1,k} - \bar{X}_{0,k}$, normalized by the square root of the sum of the sample variances for treated and controls, $\sqrt{S_{0,k}^2 + S_{1,k}^2}$ (with the latter calculated on the full sample for the first column and on the selected sample for the third column). Based on the subclassification with two or five subclasses, we also present, in columns 4 and 5,

$$
nor - dif = \sum_{j=1}^{J} g_j \cdot (\bar{X}_{1j,k} - \bar{X}_{0j,k}) / \sqrt{S_{0,k}^2 + S_{1,k}^2},
$$

(normalized by the same function of the standard deviations in the selected sample so that the normalized differences are directly comparable to those in column 3). The ratio of the fourth and fifth column to the third column shows how much the subclassification reduces the bias coming from linear effects of the covariates in the trimmed sample, that is, the $\gamma_k$ in equation (17.1). We see that those covariates that exhibit substantial differences between treated and controls in the full sample, show much less difference after trimming and even less after subsequent subclassifications. For example, consider the covariate # tickets. In the full sample there is a normalized difference of 0.64. Trimming the sample reduces that to 0.34. Subclassification with two blocks reduces that further to 0.12, and five subclasses reduces this to a paltry 0.04, or about 6% of the original 0.64. For the covariate with the second biggest normalized difference in the full sample, education, which exhibited a normalized difference of 50% of the standard deviation in the full sample, we similarly get a reduction to about 7% of the standard deviation in the trimmed sample. For covariates where the differences were originally smaller, the reduction is much less, but with the five subclasses, there are now no covariates with a normalized difference larger than 0.10, and only a single one, (namely, male), with a normalized difference in average covariates equal to 0.10. Subclassification has clearly been effective in removing the bulk of the differences in covariates for all eighteen covariates in this data set.

### 17.5 Estimation Based on Subclassification

Given the stratum definitions, we first estimate the average effect of the treatment within each stratum. The first estimator we consider, $\hat{\tau}_j$, is simply the difference between average
outcomes for treated and controls in stratum $j$:

$$
\hat{\tau}_j = \bar{Y}_{1j} - \bar{Y}_{0j} = \frac{1}{N_{1j} \sum_{i: W_i = 1}} B_{ij} \cdot Y_i - \frac{1}{N_{0j} \sum_{i: W_i = 0}} B_{ij} \cdot Y_i.
$$

The subclassification or blocking estimator for the average treatment effect is

$$
\hat{\tau}_{\text{block}} = \sum_{j=1}^{J} q_j \cdot \hat{\tau}_j.
$$

This simple estimator may not be very attractive. Even if the propensity score is known, the differences $\bar{Y}_{1j} - \bar{Y}_{0j}$ will be biased for the average treatment effects within the blocks because the propensity score is only approximately constant within the blocks. We therefore may wish to further reduce any remaining bias by modifying the basic estimator. Two leading alternatives are to use regression (covariance) adjustment or model-based imputation within the blocks. This raises an important issue regarding the choice of blocks. With many blocks, typically some will contain relatively few units, and so it may be difficult to estimate linear regression functions precisely. Therefore, if one intends to combine the subclassification with regression or model-based adjustment, one may wish to ensure a relatively large number of units in each stratum.

Here we further develop the regression approach. It is useful to start by reinterpreting the within-block difference in average treatment and control outcomes $\hat{\tau}_j$ as the least squares estimator of $\tau_j$ in the regression function

$$
Y_i^{\text{obs}} = \alpha_j + \tau_j \cdot T_i + \varepsilon_i,
$$

where as before, $T_i = 1_{W_i = 1}$. We estimate the parameters of the regression function using only the $N_j$ observations in the $j$-th stratum or block. We can then generalize this estimator to allow for covariates by specifying within block $j$ the regression function

$$
Y_i^{\text{obs}} = \alpha_j + \tau_j \cdot T_i + \beta_j' X_i + \varepsilon_i,
$$

again using only the $N_j$ observations in block $j$. If the balancing on the propensity score was perfect, the population correlation between the covariates and the treatment indicator within a block would be zero. In that case the inclusion of the covariates in this regression only serves to improve precision, the same way using covariates in the analysis of a completely randomized experiment can improve precision: on average the estimator based on (17.2)
would be the same as the estimator based on (17.3). If, however, the balancing on the
propensity score is less than perfect, the role of the regression is twofold. In addition to
improving precision it also helps in reducing any remaining bias. It is important to note that
categorically the use of regression adjustment is quite different here from using regression
methods on the full sample. Within the blocks there is less concern about using the regression
function to extrapolate out of sample, because the blocking has already ensured that the
covariate distributions are similar. In practice at this stage the use of regression methods is
more like that in randomized experiments where the similarity of the covariate distributions
greatly reduces the sensitivity to the specification of the regression function.

Mechanically the analysis now estimates the average treatment effects within the blocks
using linear regression:

\[
\left(\hat{\alpha}_j, \hat{\tau}_{adj,j}, \hat{\beta}_j\right) = \arg\min_{\alpha,\tau,\beta} \sum_{i=1}^{N_j} B_{ij} \cdot \left(Y_{i}^{obs} - \alpha - \tau \cdot T_i - \beta'X_i\right)^2,
\]

using the \(N_j\) units within stratum \(j\). Within each block the procedure is the same as that
for analyzing completely randomized experiments with regression adjustment discussed in
Chapter 6. These within-block least squares estimates \(\hat{\tau}_j\) are then again averaged to get an
 estimator for the overall average treatment effects,

\[
\hat{\tau}_{block-adj} = \sum_{j=1}^{J} q_j \cdot \hat{\tau}_{adj,j},
\]

with the stratum weights still equal to the stratum shares \(q_j = N_j/N\).

## 17.6 Inference

For the simple subclassification estimator with no further covariance adjustment we directly
apply the Neyman analysis for completely randomized experiments. Using the results from
Chapter 9 on Neyman’s repeated sampling perspective, applied in the context of stratified
randomized experiments, the variance of \(\hat{\tau}_j\) is

\[
\mathbb{V}(\hat{\tau}_j) = \frac{S_{0j}^2}{N_{0j}} + \frac{S_{1j}^2}{N_{1j}} - \frac{S_{01j}^2}{N_j},
\]

where, as in Chapter 9, the chapter on stratified randomized experiments,

\[
S_{0j}^2 = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} B_{ij} \cdot (Y_i(0) - \overline{Y_j}(0))^2, \quad S_{1j}^2 = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} B_{ij} \cdot (Y_i(1) - \overline{Y_j}(1))^2,
\]
\[ S_{01j}^2 = \frac{1}{N-1} \sum_{i=1}^{N} B_{ij} \cdot (Y_i(1) - Y_i(0) - \tau_j)^2 \]

and

\[ \bar{Y}_j(w) = \frac{1}{N_j} \sum_{i=1}^{N} B_{ij} \cdot Y_i(w). \]

To get a conservative estimate of the variance \( \mathbb{V}(\hat{\tau}_j) \) we substitute

\[ s_{0j}^2 = \frac{1}{N_{0j} - 1} \sum_{i:W_i=0} B_{ij} \cdot (Y_{i,\text{obs}} - \bar{Y}_0j)^2, \quad \text{and} \quad s_{1j}^2 = \frac{1}{N_{1j} - 1} \sum_{i:W_i=1} B_{ij} \cdot (Y_{i,\text{obs}} - \bar{Y}_1j)^2, \]

for \( s_{0j}^2 \) and \( s_{1j}^2 \) respectively, and \( s_{01j}^2 = 0 \) for \( S_{01j}^2 \) to get the following estimator,

\[ \hat{\mathbb{V}}(\hat{\tau}_j) = \frac{1}{N_{0j} \cdot (N_{0j} - 1)} \sum_{i:W_i=0} B_{ij} \cdot (Y_{i,\text{obs}} - \bar{Y}_0j)^2 + \frac{1}{N_{1j} \cdot (N_{1j} - 1)} \sum_{i:W_i=1} B_{ij} \cdot (Y_{i,\text{obs}} - \bar{Y}_1j)^2. \]

Because conditional on \( X \), the within-stratum estimator \( \hat{\tau}_j \) is independent of \( \hat{\tau}_k \) if \( j \neq k \) we can calculate the variance of \( \hat{\tau}_{\text{block}} = \sum_{j=1}^{J} q_j \cdot \hat{\tau}_j \) by adding the within-block variances, multiplied by the square of the block proportions:

\[ \hat{\mathbb{V}}(\hat{\tau}_{\text{block}}) = \sum_{j=1}^{J} \hat{\mathbb{V}}(\hat{\tau}_j) \cdot q_j^2 = \sum_{j=1}^{J} \hat{\mathbb{V}}(\hat{\tau}_j) \cdot \left( \frac{N_j}{N} \right)^2. \]

In practice, however, we typically do further covariance adjustment to reduce the remaining bias. Here we focus on the specific estimator discussed in the previous subsection, where we use linear regression within the blocks, with identical slopes in the treatment and control subsamples. We use the standard variance for ols estimators, allowing for general heteroskedasticity. Let \( \left( \hat{\alpha}_j, \hat{\tau}_{\text{adj},j}, \hat{\beta}_j \right) \) be the ordinary least squares estimates defined in equation (17.4). Then define the matrices \( \hat{\Delta} \) and \( \hat{\Gamma} \) as

\[ \hat{\Delta}_j = \frac{1}{N_j} \sum_{i=1}^{N} B_{ij} \left( \begin{array}{ccc} 1 & T_i & X'_i \\ T_i & T_i & T_i \cdot X'_i \\ X_i & T_i \cdot X_i & X_i \cdot X_i \end{array} \right), \]

and

\[ \hat{\Gamma}_j = \frac{1}{N_j} \sum_{i=1}^{N} B_{ij} \left( Y_i - \hat{\alpha}_j - \hat{\tau}_{\text{adj},j} T_i - \hat{\beta}_j X_i \right)^2 \cdot \left( \begin{array}{ccc} 1 & T_i & X'_i \\ T_i & T_i & T_i \cdot X'_i \\ X_i & T_i \cdot X_i & X_i \cdot X_i \end{array} \right). \]
Then the robust estimator for the variance of $\hat{\tau}_{adj,j}$ is $\hat{V}(\hat{\tau}_{adj,j})$, the natural generalization of the Neyman variance estimator, is

$$\hat{V}(\hat{\tau}_{adj,j}) = \frac{1}{N_j} \left( \hat{\Gamma}_j \hat{\Delta}_j^{-1} \hat{\Gamma}_j \right)^{-1}_{(2,2)},$$

the $(2,2)$ element of the $(K + 2) \times (K + 2)$ dimensional matrix $\left( \hat{\Gamma}_j \hat{\Delta}_j^{-1} \hat{\Gamma}_j \right)^{-1} / N_j$. We then combine the within-block variances the same way we did before:

$$\hat{V}(\hat{\tau}_{block-adj}) = \sum_{j=1}^{J} \hat{V}(\hat{\tau}_{adj,j}) \cdot q_j^2. \quad (17.5)$$

This is the variance we use in the calculations in the next section.

If we are interested in the average treatment effect for the treated subsample, we do not need to modify the within-block estimates $\hat{\tau}_{adj,j}$ or variances $\hat{V}(\hat{\tau}_{adj,j})$. Because we analyze the data within the blocks as if assignment is completely random, the average effect for the treated within the block is identical to the overall average effect. In order to get the average effect for the treated for the entire sample, however, we do modify the weights to reflect the differences in proportions of treated units in the different blocks. Instead of using the sample proportions $q_j = N_j / N$, the appropriate weights are now equal to the proportion of treated units in each block, $N_{1j} / N_t$, leading to

$$\hat{\tau}_{block-adj,treated} = \sum_{j=1}^{J} \hat{\tau}_{adj,j} \cdot \frac{N_{1j}}{N_t}.$$

Similarly for the variance we sum the within-block variances, multiplied by the square of the block proportions of treated:

$$\hat{V}(\hat{\tau}_{block-adj,treated}) = \sum_{j=1}^{J} \hat{V}(\hat{\tau}_{adj,j}) \cdot \left( \frac{N_{1j}}{N_t} \right)^2.$$

### 17.7 Average Treatment Effects for the Lottery Data

Now let us return to the lottery data. The algorithm for choosing the number of blocks led to five blocks. Within each of these five blocks we estimate the average treatment effect $(i)$ using either no further adjustment, $(ii)$, using linear regression with a limited number of four covariates (the same four that are always included in the specification of
the propensity score, \# tickets, education, working then, and earnings year -1), and (iii), using linear regression with the full set of eighteen covariates.

In Table 6 we present results for the parameter estimates from the least squares regression for the five blocks. We present both the least squares estimates for the case with no covariates and for the case with the limited set of four covariates. Although the parameter estimates are of only limited interest here, we note that we see some evidence that the covariates do affect the outcomes, and also that there is sufficient difference in the covariate distributions within the blocks that the adjustment makes a difference to the estimates of the effect of the treatment within the blocks.

In Table 7 we present the estimates of the overall treatment effect using subclassification. We report estimates based on the full sample, the selected sample with no subclassification, the selected sample with two blocks, and the selected sample with the optimal number of five blocks. In each case we present the estimates based on no further covariance adjustment, covariance adjustment with the limited set of four covariates, and covariance adjustment based on the full set of eighteen covariates. The key insight is that both trimming and subclassification greatly reduce the sensitivity to the inclusion of covariates in the regression specification. In the full sample, the estimates range from -6.2 to -2.8 (in terms of thousands of dollars) in reduced labor earnings as a result of winning the lottery, a range of 3.4. In the trimmed sample the estimates range from -6.6 to -4.0, a range of 2.6. In the trimmed sample with two blocks the range is only 0.8, and in the case with five blocks the range is down to 0.6. The conclusion is that, at least for this data set, subclassification greatly reduces the sensitivity to the specific methods used, and thus leads to more robust estimates of the causal effects of interest.

### 17.8 Weighting Estimators and Subclassification

There is an alternative way of using the propensity score that is at first sight very different from, although in reality fairly closely related to subclassification. In this approach, ultimately derived from the work by Horvitz and Thompson (1952) in survey research, the propensity score is used to weight the units in order to eliminate biases associated with differences in observed covariates.
17.8.1 Weighting Estimators

The Horvitz-Thompson estimator exploits the following two equalities, which follow from unconfoundedness:

\[
\mathbb{E}
\left[
\frac{T_i \cdot Y_i^{obs}}{e(X_i)}
\right]
= \mathbb{E}[Y_i(1)]
\quad \text{and} \quad
\mathbb{E}
\left[
\frac{(1 - T_i) \cdot Y_i^{obs}}{1 - e(X_i)}
\right]
= \mathbb{E}[Y_i(0)].
\tag{17.6}
\]

These inequalities can be derived as follows. Because \(Y_i^{obs}\) is \(Y_i(1)\) if \(W_i = 1\) (or \(T_i = 1\)),

\[
\mathbb{E}
\left[
\frac{T_i \cdot Y_i^{obs}}{e(X_i)}
\right]
= \mathbb{E}
\left[
\frac{T_i \cdot Y_i(1)}{e(X_i)}
\right].
\]

By iterated expectations, we can write this as

\[
\mathbb{E}
\left[
\frac{T_i \cdot Y_i(1)}{e(X_i)} \mid X_i
\right]
= \mathbb{E}
\left[
\frac{T_i}{e(X_i)} \cdot \mathbb{E}[Y_i(1) \mid X_i]
\right] = \mu_1(X_i),
\]

and thus

\[
\mathbb{E}
\left[
\frac{T_i \cdot Y_i(1)}{e(X_i)}
\right]
= \mathbb{E}[\mu_1(X_i)] = \mathbb{E}[Y_i(1)].
\]

Essentially the same argument leads to the second equality for the average control outcome.

The two equalities in (17.6) suggest estimating \(\mathbb{E}[Y_i(1)]\) and \(\mathbb{E}[Y_i(0)]\) as

\[
\overline{Y_i(1)} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i \cdot Y_i^{obs}}{e(X_i)}
\quad \text{and} \quad
\overline{Y_i(0)} = \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - T_i) \cdot Y_i^{obs}}{1 - e(X_i)},
\]

and thus estimating the average treatment effect \(\tau = \mathbb{E}[Y_i(1) - Y_i(0)]\) as

\[
\bar{\tau}_{ht} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{T_i \cdot Y_i^{obs}}{e(X_i)} - \frac{(1 - T_i) \cdot Y_i^{obs}}{1 - e(X_i)}\right)
\quad \text{and} \quad
\bar{\tau}_{ht} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(T_i - e(X_i)) \cdot Y_i^{obs}}{e(X_i) \cdot (1 - e(X_i))}\right),
\tag{17.7}
\]

We can think of this estimator as a weighted average of the outcomes,

\[
\bar{\tau}_{ht} = \frac{1}{N} \sum_{i=1}^{N} (2 \cdot T_i - 1) \cdot Y_i^{obs} \cdot \tilde{\lambda}_i,
\]
with weights:

$$\tilde{\lambda}_{ht,i} = \begin{cases} \frac{1}{1 - \hat{e}(X_i)} & \text{if } W_i = 0, \\ \frac{1}{\hat{e}(X_i)} & \text{if } W_i = 1. \end{cases}$$

The weights for both the treated and control units add up in expectation to $N$,

$$E \left[ \sum_{i: W_i = 0} \lambda_i \right] = N, \quad \text{and} \quad E \left[ \sum_{i: W_i = 1} \lambda_i \right] = N,$$

so that the marginal expectation of the weights is equal to two.

In practice we do not know the propensity score, so we cannot use the estimator in (17.7) directly. Instead we weight using the estimated propensity score $\hat{e}(X_i)$, and use the estimator

$$\hat{\tau}_{ht} = \frac{\sum_{i=1}^{N} T_i \cdot Y_{i}^{\text{obs}}}{\sum_{i=1}^{N} \hat{e}(X_i)} - \frac{\sum_{i=1}^{N} (1 - T_i) \cdot Y_{i}^{\text{obs}}}{\sum_{i=1}^{N} 1 - \hat{e}(X_i)} - \frac{\sum_{i=1}^{N} T_i}{\sum_{i=1}^{N} 1 - \hat{e}(X_i)}.$$  \hfill (17.8)

(Normalizing the weights to add up to one improves the properties of the estimator slightly.)

The basic Horvitz-Thompson estimator can be modified easily to incorporate covariates. For this purpose it is useful to write the weighting estimator as a weighted regression estimator. Consider the regression function

$$Y_{i}^{\text{obs}} = \alpha + \tau \cdot T_i + \varepsilon_i,$$

estimated by weighted least squares with weights $\lambda_{ht,i}$, where

$$\lambda_{ht,i} = \begin{cases} \frac{1}{1 - \hat{e}(X_i)} & \text{if } W_i = 0, \\ \frac{1}{\hat{e}(X_i)} & \text{if } W_i = 1. \end{cases}$$

(The only difference between $\lambda_{ht,i}$ and $\tilde{\lambda}_{ht,i}$ is that the former depends on the estimated propensity scores $\hat{e}(X_i)$, and the latter on the true propensity scores $e(X_i)$.) This weighted regression estimator for $\tau$ is identical to $\hat{\tau}_{ht}$ as defined in (17.8). In this weighted regression version it is straightforward to include covariates. Instead of estimating the regression function with only an intercept and an indicator for the treatment, one can estimate a regression function that includes additional covariates:

$$Y_{i}^{\text{obs}} = \alpha + \tau \cdot T_i + \beta'X_i + \varepsilon_i,$$

using the same weights $\lambda_{ht,i}$. This estimator that combines weighting with regression has been developed by Jamie Robins and coauthors (e.g., Robins, Zhao and Rotnitzky, 1995). They
show that the weighted regression estimator is consistent as long as either the specification of the propensity score is correct, or the specification of the regression function is correct, a property Robins and coauthors refer to as “double-robustness.” We discuss this in more detail in the appendix to this chapter.

It is useful to see how this Horvitz-Thompson estimator relates to the subclassification estimator. The basic subclassification estimator, with no further adjustment for covariates, has the form

$$
\hat{\tau}_{\text{block}} = \sum_{j=1}^{J} q_j \cdot \hat{\tau}_j = \sum_{j=1}^{J} q_j \cdot (Y_{1j} - Y_{0j}),
$$

which can be written as

$$
\hat{\tau}_{\text{block}} = \frac{1}{N} \sum_{i=1}^{N} T_i \cdot Y_{\text{obs}} \cdot \lambda_i - \frac{1}{N} \sum_{i=1}^{N} (1 - T_i) \cdot Y_{\text{obs}} \cdot \lambda_{\text{block},i},
$$

where the weights $\lambda_{\text{block},i}$ satisfy

$$
\lambda_{\text{block},i} = \begin{cases} 
\frac{1}{N_{0j}/N_j} & \text{if } W_i = 0, B_{ij} = 1, \\
\frac{1}{N_{1j}/N_j} & \text{if } W_i = 1, B_{ij} = 1.
\end{cases}
$$

Thus the basic subclassification estimator can be interpreted as a weighting estimator where the weights are based on the adjusted propensity score. Instead of using the original estimator for the propensity score, $\hat{e}(X_i)$, the blocking estimator uses the fraction of treated units within the propensity score stratum that the unit belongs to:

$$
\tilde{e}(X_i) = \sum_{j=1}^{J} B_{ij} \cdot \frac{N_{1j}}{N_j}.
$$

Thus, it coarsens the propensity score, essentially averaging it within the subclasses. This increases very low values of the propensity score, and decreases very high values, and thus lowers extreme values for the weights $\lambda_i$ in the weighted average interpretation of the estimator.

What are the relative merits of the subclassification estimator versus the Horvitz-Thompson estimator? We discuss four issues. In the end we prefer the subclassification estimator and see little reason to use the estimator based on weighting on the estimated propensity score. First of all, we should note that in many cases this choice is not important, as it will not make much difference whether one uses the Horvitz-Thompson or subclassification weights. If the
number of blocks is large, so that the dispersion of the propensity score within the strata is limited, then the weights according to the blocking estimator will be close to those according to the Horvitz-Thompson estimator. This is also true if there is only limited variation in the propensity score overall, and if there are few extreme values for the propensity score. The weights will be different only if, in at least some blocks, there is substantial variation in the propensity score. This is most likely to happen in blocks with propensity score values close to zero and one. In fact, the similarity between the estimators turns to equality in simple cases where the model for the propensity score is fully saturated and the number of blocks is sufficiently large that within a block there is no variation at all in the propensity score.

A second point concerns bias properties of the two estimators. If one uses the true propensity score, the Horvitz-Thompson estimator is exactly unbiased. If one does not know the propensity score, it might then seem that using the best possible estimate of the propensity score, rather than one that is further smoothed, may be sensible. This appears the most powerful argument in favor of the Horvitz-Thompson estimator. This argument may not be particularly persuasive though. Although weighting by the true propensity score leads to unbiased estimators for the average treatment effect, weighting using a noisy, even if unbiased estimator for the propensity score may have considerable bias given that the propensity score enters in the denominator of the weights. Smoothing the weights by essentially averaging them within blocks as the subclassification estimator does, may remove some of this bias. Moreover, in practice the propensity score is likely to be misspecified. This may affect the performance of the Horvitz-Thompson estimator more than the subclassification estimator. More specifically, suppose a particular covariate \( X_{i,k} \) is omitted from the propensity score specification. If this covariate is correlated with the potential outcomes, any (small) bias from omitting it may be increased by the larger weights used in the Horvitz-Thompson approach.

The third point concerns the variance. Here it the relative merits are clear. By smoothing over the extreme weights from the Horvitz-Thompson estimator, the subclassification estimator tends to have a smaller variance. This may make also the Horvitz-Thompson estimator less robust than the blocking estimator as the large weights also tend to be the ones that are relatively imprecisely estimated. For that reason, shrinking them to their mean, as subclassification does, can improve the properties of the resulting estimator.

A fourth issue concerns the modifications of the Horvitz-Thompson and blocking estima-
tor involving additional covariance adjustment. The covariance adjustment version of the Horvitz-Thompson estimator uses a single set of parameters to model the dependence of the outcome on the covariates. In other words, it uses a global approximation to the regression function. Such a global approximation can lead to poor approximations for some values of the covariates. The analogous procedure given the subclassification would be to restrict the slope coefficients on the covariates to be the same in all blocks. This is not what is typically, or what we discussed in the previous sections. Instead the slope coefficients are unrestricted between the blocks, allowing the regression functions to provide better approximations to the regression functions for each subclass.

17.8.2 Weighting Estimators and the Lottery Data

To illustrate the Horvitz-Thompson estimator let us go back to the lottery data. We look both at the full sample with 496 units and at the trimmed sample with 323 units. In both cases we calculate the weights according to the propensity score estimated through the algorithm discussed before. Based on the estimated propensity score we normalize the weights within each treatment group to ensure they add up to $N$. We then estimate the implicit weights in the blocking estimator, again for both the full sample and for the trimmed sample.

Table 8 presents summary statistics for the weights. Within each data set there is a fair amount of difference between the Horvitz-Thompson and subclassification weights. In the full sample the correlation coefficient between the Horvitz-Thompson and subclassification weights is only 0.64. In the trimmed sample the correlation is considerably higher, namely 0.82. The second insight is that the weights are considerably more extreme for the Horvitz-Thompson estimator. In the full sample the largest of the weights is almost 80 for the Horvitz-Thompson estimator, compared to 17.8 for the subclassification estimator. With the smallest weights around one (the smallest weight would be at least equal to one if it was not for the normalization to ensure that the weights add up to the sample size), the weight for this unit is eighty times that for the low weight unit, making any estimates very sensitive to the outcome for this unit. Increasing the outcome for this individual by one standard deviation (that is, increasing average post lottery earnings by fifteen thousand dollars), would lead to a change in the estimated average treatment effect of $(80/496) \times 15 = 2.5$. 
This is a substantial amount, given the variation in our subclassification estimates in Table 7. The sensitivity of the estimates to the outcome for this unit in the subclassification estimator is only a fifth of that. In the trimmed sample the largest weights are 18.2 and 6.2 for the Horvitz-Thompson and subclassification estimators, so now changing the outcome for any single unit by a standard deviation leads to a change in the estimated average effect of at most \((6.2/323) \times 15 = 0.3\). In particular for the subclassification in the trimmed sample the ratio of largest to smallest weight is 5.2, ensuring that no single units is unduly affecting the estimates. The third insight is that the trimming greatly reduces the variation in the weights, and lowers the highest weights, by improving the balance and shrinking the propensity score towards average values. In general the subclassification smoothes out the weights, avoiding excessively large weights.

Let us pursue the relative merits of subclassification and propensity score weighting a little further here. Suppose, as we have done before in illustrative calculations, that the conditional expectation of the potential outcomes is linear in the covariates:

\[
E\left[ Y_i(w) \mid X_i = x \right] = \alpha + \tau \cdot T_i + \beta'x,
\]

with constant variance:

\[
\mathbb{V} \left( Y_i(w) \mid X_i = x \right) = \sigma^2.
\]

If this linearity assumption were actually true, we could simply estimate \( \tau \) by least squares. However, the point is not getting an estimate of the average treatment effect under this assumption, but rather to assess the properties of the two propensity-score based estimators, the Horvitz-Thompson estimator versus the subclassification estimator, under this assumption.

An estimator of the form

\[
\hat{\tau}_\lambda = \frac{1}{N} \sum_{i=1}^{N} \left( T_i \cdot Y_i^{\text{obs}} \cdot \lambda_i - (1 - T_i) \cdot Y_i^{\text{obs}} \cdot \lambda_i \right),
\]

with the weights \( \lambda_i \) of the form \( \lambda_i = \lambda(W_i, X_i, W_{(i)}, X_{(i)}) \), has, conditional on \( W \) and \( X \), the following bias and variance:

\[
\text{Bias}_\lambda = E \left[ \hat{\tau}_\lambda - \tau \mid W, X \right] = \frac{1}{N} \sum_{i=1}^{N} \left( T_i \cdot \mu_1(X_i) \cdot \lambda_i - (1 - T_i) \cdot \mu_0(X_i) \cdot \lambda_i \right) - \tau,
\]
and

$$\text{Var}_\lambda = \mathbb{V}(\hat{\tau}_\lambda | \bf{W}, \bf{X}) = \frac{1}{N^2} \sum_{i=1}^{N} \lambda_i^2 \cdot \sigma_{\hat{W}_i}^2(X_i).$$

Under our linear model assumptions the bias simplifies to

$$\text{Bias}_\lambda = \beta' \frac{1}{N} \sum_{i=1}^{N} (2 \cdot T_i - 1) \cdot X_i \cdot \lambda_i = \beta' (\overline{X}_{\text{1, weighted}} - \overline{X}_{\text{0, weighted}}),$$

where

$$\overline{X}_{\text{0, weighted}} = \sum_{i: W_i = 0} X_i \cdot \lambda_i / \sum_{i: W_i = 0} \lambda_i, \quad \text{and} \quad \overline{X}_{\text{1, weighted}} = \sum_{i: W_i = 1} X_i \cdot \lambda_i / \sum_{i: W_i = 1} \lambda_i,$$

the weighted average of the control and treated covariates. Under homoskedasticity the variance simplifies to

$$\text{Var}_\lambda = \frac{\sigma^2}{N^2} \cdot \sum_{i=1}^{N} \lambda_i^2.$$  

We can estimate these two objects, Bias\(_\lambda\) and Var\(_\lambda\), and the sum of the variance and the square of the bias, that is, the expected-mean-squared-error, for the particular data set. The results are reported in Tables 6 and 10. In Table 6 we report the least squares estimates of the regression function, for both the full and the trimmed sample. In Table 1 in columns 2 and 6 we report the average difference in covariates, weighted by according to the Horvitz-Thompson estimator, and normalized by the square root of the sum of the standard deviations

$$\frac{\overline{X}_{\text{1, weighted}} - \overline{X}_{\text{0, weighted}}}{\sqrt{S_0^2 + S_1^2}},$$

If the Horvitz-Thompson estimator was based on the true propensity scores, the average difference in covariates should be zero, at least in expectation. It is not, due in part to sampling variation and due in part to misspecification of the propensity score. We see that for most covariates the Horvitz-Thompson estimator has approximately the same normalized differences as the subclassification estimator. Sometimes they are higher, as for example, for the important (in the sense of being likely to be correlated with the potential outcomes) lagged earnings variables, sometimes smaller, as for education and some of the employment
indicators. The higher normalized differences are largely due to the presence of large weights in the Horvitz-Thompson approach.

In Table 10 we report the components of the expected-mean-squared-error. Unsurprisingly we find that for both the full and the trimmed sample the variance is lower for the subclassification estimator. This is a direct consequence of the coarsening of the subclassification estimator. It smooths out the extreme weights. More surprising is the fact that for both the full sample, and the trimmed sample, the bias is actually considerably larger for the Horvitz-Thompson estimator than for the subclassification estimator. Not surprisingly the bias and variance are substantially smaller in the trimmed sample than in the full sample (with the exception of the variance for the subclassification estimator which is slightly smaller in the full sample than in the trimmed sample).

17.9 Conclusion

In this chapter we discuss one of the leading estimators for average treatment effects under unconfoundedness. This subclassification estimator uses the propensity score to construct strata within which the covariates are well balanced. Within the strata the average treatment effect is estimated by simply differencing average outcomes for treated and control units, or, in our preferred version, by further adjusting for remaining covariate differences through regression. This estimator is similar to weighting estimators, although less variable in settings with units with propensity score values close to zero or one. We illustrate the practical value of this estimator using the lottery data.
Appendix: Variance Calculations

Rosenbaum and Rubin
Notes

Imbens (multiple treatments)
Hirano, Imbens and Ridder
### Table 1: Normalized Differences in Covariates after Subclassification

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full Sample</th>
<th></th>
<th>Trimmed Sample</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 Block</td>
<td>HT</td>
<td>1 Block</td>
<td>2 Blocks</td>
<td>5 Blocks</td>
<td>HT</td>
<td></td>
</tr>
<tr>
<td>Year Won</td>
<td>-0.19</td>
<td>0.07</td>
<td>-0.04</td>
<td>-0.02</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Tickets</td>
<td>0.64</td>
<td>0.07</td>
<td>0.36</td>
<td>0.12</td>
<td>0.05</td>
<td>-0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.33</td>
<td>-0.21</td>
<td>-0.06</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>-0.13</td>
<td>0.06</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.10</td>
<td>-0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>-0.50</td>
<td>0.31</td>
<td>-0.34</td>
<td>-0.13</td>
<td>-0.07</td>
<td>-0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work Then</td>
<td>0.06</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earn Year -6</td>
<td>-0.19</td>
<td>0.01</td>
<td>-0.13</td>
<td>-0.07</td>
<td>-0.02</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earn Year -5</td>
<td>-0.20</td>
<td>0.01</td>
<td>-0.14</td>
<td>-0.05</td>
<td>-0.00</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earn Year -4</td>
<td>-0.21</td>
<td>-0.01</td>
<td>-0.16</td>
<td>-0.06</td>
<td>-0.01</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earn Year -3</td>
<td>-0.18</td>
<td>0.03</td>
<td>-0.13</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earn Year -2</td>
<td>-0.19</td>
<td>0.04</td>
<td>-0.14</td>
<td>-0.02</td>
<td>0.01</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earn Year -1</td>
<td>-0.16</td>
<td>0.08</td>
<td>-0.13</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos Earn Year -6</td>
<td>0.02</td>
<td>0.11</td>
<td>-0.00</td>
<td>-0.06</td>
<td>-0.06</td>
<td>-0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos Earn Year -5</td>
<td>0.10</td>
<td>-0.10</td>
<td>0.07</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos Earn Year -4</td>
<td>0.07</td>
<td>-0.14</td>
<td>0.04</td>
<td>-0.00</td>
<td>-0.01</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos Earn Year -3</td>
<td>0.09</td>
<td>-0.13</td>
<td>0.02</td>
<td>-0.03</td>
<td>-0.04</td>
<td>-0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos Earn Year -2</td>
<td>0.10</td>
<td>-0.12</td>
<td>0.04</td>
<td>0.00</td>
<td>-0.03</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos Earn Year -1</td>
<td>0.07</td>
<td>0.12</td>
<td>-0.01</td>
<td>-0.03</td>
<td>-0.05</td>
<td>-0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Sample Sizes for Selected Subsamples with the Propensity Score between $\alpha$ and $1 - \alpha$ (Estimated $\alpha = 0.0891$).

<table>
<thead>
<tr>
<th></th>
<th>low $\hat{e}(X_i) &lt; \alpha$</th>
<th>middle $\alpha \leq \hat{e}(X_i) \leq 1 - \alpha$</th>
<th>high $1 - \alpha &lt; \hat{e}(X_i)$</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Losers</td>
<td>82</td>
<td>172</td>
<td>5</td>
<td>259</td>
</tr>
<tr>
<td>Winners</td>
<td>4</td>
<td>151</td>
<td>82</td>
<td>237</td>
</tr>
<tr>
<td>All</td>
<td>86</td>
<td>323</td>
<td>87</td>
<td>496</td>
</tr>
</tbody>
</table>

Table 3: Estimates of Propensity Score in Selected Sample

<table>
<thead>
<tr>
<th>Covariate</th>
<th>intercept</th>
<th>21.77</th>
<th>0.13</th>
<th>164.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear terms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Tickets</td>
<td>-0.08</td>
<td>0.46</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>-0.45</td>
<td>0.08</td>
<td>-5.7</td>
<td></td>
</tr>
<tr>
<td>Working Then</td>
<td>3.32</td>
<td>1.95</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Earnings Year -1</td>
<td>-0.02</td>
<td>0.01</td>
<td>-1.4</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.05</td>
<td>0.01</td>
<td>-3.7</td>
<td></td>
</tr>
<tr>
<td>Pos Earnings Year -5</td>
<td>1.27</td>
<td>0.42</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Year Won</td>
<td>-4.84</td>
<td>1.53</td>
<td>-3.2</td>
<td></td>
</tr>
<tr>
<td>Earnings Year -5</td>
<td>-0.04</td>
<td>0.02</td>
<td>-2.1</td>
<td></td>
</tr>
<tr>
<td>quadratic terms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year Won × Year Won</td>
<td>0.37</td>
<td>0.12</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Tickets Bought × Year Won</td>
<td>0.14</td>
<td>0.06</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>Tickets Bought × Tickets Bought</td>
<td>-0.04</td>
<td>0.02</td>
<td>-1.8</td>
<td></td>
</tr>
<tr>
<td>Working Then × Year Won</td>
<td>-0.49</td>
<td>0.30</td>
<td>-1.6</td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Optimal Subclassification

<table>
<thead>
<tr>
<th>Subclass</th>
<th>Min P-score</th>
<th>Max P-score</th>
<th># Controls</th>
<th># Treated</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
<td>0.24</td>
<td>67</td>
<td>13</td>
<td>-0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.24</td>
<td>0.32</td>
<td>32</td>
<td>8</td>
<td>0.9</td>
</tr>
<tr>
<td>3</td>
<td>0.32</td>
<td>0.44</td>
<td>24</td>
<td>17</td>
<td>1.7</td>
</tr>
<tr>
<td>4</td>
<td>0.44</td>
<td>0.69</td>
<td>34</td>
<td>47</td>
<td>2.0</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
<td>0.99</td>
<td>15</td>
<td>66</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Table 5: Subclassification with Two Subclasses

<table>
<thead>
<tr>
<th>Subclass</th>
<th>Min P-score</th>
<th>Max P-score</th>
<th># Controls</th>
<th># Treated</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
<td>0.44</td>
<td>123</td>
<td>38</td>
<td>2.8</td>
</tr>
<tr>
<td>2</td>
<td>0.44</td>
<td>0.99</td>
<td>49</td>
<td>113</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 6: Least Squares Regressions Within Blocks

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Block 1 (N=80)</th>
<th>Block 2 (N=40)</th>
<th>Block 3 (N=41)</th>
<th>Block 4 (N=81)</th>
<th>Block 5 (N=81)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>est</td>
<td>s.e.</td>
<td>est</td>
<td>s.e.</td>
<td>est</td>
</tr>
<tr>
<td>no covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intercept</td>
<td>20.02</td>
<td>(2.25)</td>
<td>12.70</td>
<td>(2.67)</td>
<td>15.59</td>
</tr>
<tr>
<td>treatment</td>
<td>-10.82</td>
<td>(4.70)</td>
<td>2.07</td>
<td>(5.10)</td>
<td>-1.17</td>
</tr>
<tr>
<td>limited covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intercept</td>
<td>-20.04</td>
<td>(10.66)</td>
<td>4.47</td>
<td>(9.80)</td>
<td>-9.91</td>
</tr>
<tr>
<td>treatment</td>
<td>-6.21</td>
<td>(4.01)</td>
<td>-6.51</td>
<td>(3.86)</td>
<td>-4.81</td>
</tr>
<tr>
<td># Tickets</td>
<td>-3.48</td>
<td>(1.39)</td>
<td>1.17</td>
<td>(1.26)</td>
<td>1.85</td>
</tr>
<tr>
<td>Education</td>
<td>2.03</td>
<td>(0.87)</td>
<td>-0.37</td>
<td>(0.81)</td>
<td>0.48</td>
</tr>
<tr>
<td>Work Then</td>
<td>-2.66</td>
<td>(2.96)</td>
<td>-0.51</td>
<td>(1.84)</td>
<td>5.98</td>
</tr>
<tr>
<td>Earn Year -1</td>
<td>0.84</td>
<td>(0.06)</td>
<td>0.83</td>
<td>(0.09)</td>
<td>0.60</td>
</tr>
</tbody>
</table>
Table 7: AVERAGE TREATMENT EFFECTS WITH OPTIMAL SUBCLASSIFICATION

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Full Sample 1 Block</th>
<th>Selected Sample 1 Block</th>
<th>Selected Sample 2 Blocks</th>
<th>Selected Sample 5 Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>-6.2 (1.4)</td>
<td>-6.6 (1.7)</td>
<td>-6.0 (1.9)</td>
<td>-5.7 (2.0)</td>
</tr>
<tr>
<td># Tickets, Education, Work Then, Earn Year -1</td>
<td>-2.8 (0.9)</td>
<td>-4.0 (1.1)</td>
<td>-5.6 (1.2)</td>
<td>-5.1 (1.2)</td>
</tr>
<tr>
<td>All</td>
<td>-5.1 (1.0)</td>
<td>-5.3 (1.1)</td>
<td>-6.4 (1.1)</td>
<td>-5.7 (1.1)</td>
</tr>
</tbody>
</table>

Table 8: WEIGHTS FOR HORVITZ-THOMPSON AND SUBCLASSIFICATION ESTIMATORS

<table>
<thead>
<tr>
<th></th>
<th>Full Sample Horvitz-Thompson subclass</th>
<th>Trimmed Sample Horvitz-Thompson subclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>maximum</td>
<td>79.79</td>
<td>17.71</td>
</tr>
<tr>
<td>standard deviation</td>
<td>4.20</td>
<td>2.63</td>
</tr>
</tbody>
</table>
Table 9: Least Squares Regression Estimates

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Full Sample est (s.e.)</th>
<th>Trimmed Sample est (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>21.20 (4.80)</td>
<td>22.76 (6.49)</td>
</tr>
<tr>
<td>Treatment Indicator</td>
<td>-5.08 (0.95)</td>
<td>-5.34 (1.08)</td>
</tr>
<tr>
<td>Year Won</td>
<td>-0.64 (0.34)</td>
<td>-0.34 (0.44)</td>
</tr>
<tr>
<td># Tickets</td>
<td>0.06 (0.15)</td>
<td>0.31 (0.21)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.26 (0.04)</td>
<td>-0.29 (0.05)</td>
</tr>
<tr>
<td>Male</td>
<td>-0.58 (0.89)</td>
<td>0.44 (1.17)</td>
</tr>
<tr>
<td>Education</td>
<td>0.04 (0.20)</td>
<td>-0.12 (0.27)</td>
</tr>
<tr>
<td>Work Then</td>
<td>0.93 (1.12)</td>
<td>1.30 (1.45)</td>
</tr>
<tr>
<td>Earn Year -6</td>
<td>-0.00 (0.11)</td>
<td>0.01 (0.14)</td>
</tr>
<tr>
<td>Earn Year -5</td>
<td>-0.02 (0.13)</td>
<td>-0.02 (0.17)</td>
</tr>
<tr>
<td>Earn Year -4</td>
<td>0.02 (0.12)</td>
<td>0.01 (0.14)</td>
</tr>
<tr>
<td>Earn Year -3</td>
<td>0.29 (0.12)</td>
<td>0.36 (0.15)</td>
</tr>
<tr>
<td>Earn Year -2</td>
<td>0.04 (0.11)</td>
<td>-0.20 (0.16)</td>
</tr>
<tr>
<td>Earn Year -1</td>
<td>0.48 (0.08)</td>
<td>0.64 (0.11)</td>
</tr>
<tr>
<td>Pos Earn Year -6</td>
<td>0.19 (1.66)</td>
<td>0.05 (2.18)</td>
</tr>
<tr>
<td>Pos Earn Year -5</td>
<td>1.78 (2.10)</td>
<td>1.44 (2.72)</td>
</tr>
<tr>
<td>Pos Earn Year -4</td>
<td>-1.04 (1.99)</td>
<td>-0.28 (2.45)</td>
</tr>
<tr>
<td>Pos Earn Year -3</td>
<td>-1.60 (1.90)</td>
<td>-2.65 (2.50)</td>
</tr>
<tr>
<td>Pos Earn Year -2</td>
<td>-1.08 (2.01)</td>
<td>0.30 (2.98)</td>
</tr>
<tr>
<td>Pos Earn Year -1</td>
<td>-0.36 (1.79)</td>
<td>-2.52 (2.65)</td>
</tr>
</tbody>
</table>

\[ \sigma^2 \quad 8.45^2 \quad 8.59^2 \]

Table 10: Bias and Variance for Weighting and Subclassification Estimator under Linear Model

<table>
<thead>
<tr>
<th></th>
<th>Full Sample Horvitz-Thompson subclass</th>
<th>Trimmed Sample Horvitz-Thompson subclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>4.34</td>
<td>1.29</td>
</tr>
<tr>
<td>Variance</td>
<td>2.59²</td>
<td>1.29²</td>
</tr>
<tr>
<td>Bias² + Variance</td>
<td>5.06²</td>
<td>1.83²</td>
</tr>
</tbody>
</table>

\[ \sigma^2 \quad 8.45^2 \quad 8.59^2 \]