**Chapter 13**

**Propensity Scoring**

with Amr ElRafey

## **[H1]** Learning Objectives

**[INSERT NL]**

1. Verify that propensity scores balance data
2. Use propensity scores to remove confounding in the data
3. Use quintiles to match treated and untreated patients on propensity scores
4. Use inverse propensity score weights to control confounding

**[END NL]**

## [H1] Key Concepts

**[INSERT BL]**

* Confounding
* Propensity score
* Quintile matching
* Inverse probability weights
* Overlap
* Extreme weight

**[END BL]**

## [H1] Chapter at a Glance

When using observational data, it is important to remove the effects of confounding statistically. Propensity scoring is one method for doing so. Patients self-select treatment programs. When analysts examine effectiveness of treatment, it is important to control for patient characteristics that might explain the observed outcomes. In these situations, propensity scoring can help. This chapter shows how the probability of seeking treatment can be predicted from the patients’ characteristics and how this probability can then be used to balance the data so that treated and untreated patients do not differ in their characteristics. An unconfounded estimate of the effect of treatment can be assessed when treated and untreated groups do not differ in their underlying characteristics. Chapter 13 shows how one can estimate the unconfounded effectiveness of treatment.

[H1] Widespread Use

Propensity scoring is widely used. There are many examples of the use of propensity scores, including the following:

**[INSERT NL]**

1. Health plans often must set a formulary for discounted medication that their clinicians are expected to prescribe to the members. The decision to include or drop a medication requires careful head-to-head comparisons of two or more medications for treatment of the same condition. Since the patients who take these medications differ in many ways, propensity scoring is used to remove patient differences and examine the outcome of the medications independent of patients who have taken them. An example of head-to-head comparison of alternative medications can be seen at Michigan Medicine, where Scappaticci and colleagues (2018) investigated the cost and efficacy of two different medications for treatment of acute myeloid leukemia. They were able to match 32 patients who took the two different medications on all aspects except the medication they took. Therefore, the researchers were able to report the impact of the medications without concern for patient differences.
2. Managers can use propensity scoring to think through how to organize their own delivery systems. For example, many recommend the use of patient-centered medical homes (PCMHs) for chronically ill patients. Propensity scores can be used to see the impact of PCMHs on patient satisfaction and health outcomes. Marsteller and colleagues (2018) studied multipayer programs. They used propensity scoring to make sure that patients in medical homes had the same rate of comorbidities than those not in medical homes. Then they estimated the impact of medical homes on various outcomes.
3. Vendors often pitch new technologies to hospitals, and it is important to examine carefully whether the new technology is less costly and has a better outcome than what the institution may currently use. For example, robotic surgery has been developed as an alternative to laparoscopic surgery, and it has been held up as less costly, with better outcomes. However, different technologies are used for different patients. Kim and colleagues (2015) removed the patient variation and concluded that robotic surgery cost more and had similar outcomes.
4. Managers may wish to pay their contractors based on the quality of their care. Propensity scoring can clarify whether pay-for-performance schemes are working. For example, Grabowski and colleagues (2017) used propensity scores to examine the impact of paying nursing homes based on their quality. They found that the nursing homes made few changes to their operations and value-based payments had no immediate impact.
5. Managers can use propensity scoring to see which employee or group of employees is responsible for patient satisfaction. Patient satisfaction reports are presented at the unit level, and it is not always clear which staff member is responsible for the rating. Propensity scoring can help identify the contribution of individual members by balancing the contribution of other staff. Similar analysis is needed for unbundling combined payments for hospital and nursing home care. Hospital managers are increasingly paid a bundled price and they need to figure out if their current downstream contractors (typically nursing homes) are costing them too much. Propensity scoring can be used to remove patient differences and select the more cost-efficient home. An example is presented in chapter 16, where we show how data balancing can clarify the roles of team members in combined outcomes such as satisfaction ratings or bundled payments.

**[END NL]**

## [H1] Propensity Scoring Is a Simulation

In random clinical trials, the covariates are balanced through randomization so that treated and untreated groups have the same rate of occurrence of covariates. In observational studies—for example, projects using data from electronic health records—this is not the case. In observational studies, the effect of treatment and covariates are confounded, making the interpretation of statistical analysis dubious.

In 1983, statisticians Paul R. Rosenbaum and Donald B. Rubin proposed to solve this problem through a method that has come to be known as *propensity scoring*. They used the covariates to predict the probability of assignment or self-selection into treatment. This probability is referred to as *propensity to participate in treatment*. Instead of focusing the analysis on the impact of treatment on outcomes in the current data set, they simulated a new sample where the rate of occurrence of covariates among treated and untreated patients was the same. Then they examined the impact of treatment in this new simulated data.

One way to understand propensity scoring is to assume that there are just three subjects; each subject is defined by a set of covariates. One has a low propensity to join, another average, and still another a very high propensity for joining. Clearly, if we examine the impact of treatment for these three subjects, differences in propensity for seeking treatment will distort the impact. Rosenbaum and Rubin proposed methods in which differences in propensity to seek treatment can be controlled statistically.

The analyst could match treated and untreated groups so that patients with low or high propensity to participate are equally observed among treated or untreated groups. She might weight the three cases inversely to their propensity, so that the patient with low propensity to be in the treated group would count several times more than the patient with high propensity. If there are fewer patients with low propensity, the weighting will increase their frequency, so that treated and untreated groups have the same number of low-propensity patients. One could also stratify the sample by propensity to be treated and then examine the impact in each stratum—in our case, strata divided according to low, average, and high propensity to seek treatment.

No matter how we remove the effect of propensity to participate in treatment, the net result is the same: a new sample of data is simulated in which the covariates occur equally among treated and untreated groups. Although the new simulated sample changes which patients are part of the sample, it does not change the relationship between treatment and outcome for each patient. Therefore, we can estimate the unconfounded impact of treatment on outcomes in this new sample.

## [H1] Three Steps in Propensity Scoring

The goal of propensity scoring is to balance the data so that patients in and out of the program do not differ in their characteristics. Propensity scoring involves three distinct steps:

**[INSERT NL]**

1. A propensity score is created that relates patient characteristics to participation in a treatment program. Because treatment is binary, logistic regression is typically used. In fact, in chapter 12, we showed how one might predict propensity to participate in the medical foster home (MFH) program. Patient characteristics that were significant in predicting participation must be controlled for.
2. The propensity score is used to balance the data through matching, weighting, or stratification. In this chapter, we focus on matching and weighting. Later, in chapter 16, we also show how the same procedures can be carried out through stratification.
3. Once balanced data have been created, the analyst examines the impact of treatment on outcome. Note that propensity scoring never changes the relationship of treatment to outcome but does change who has received treatment. Moreover, propensity scoring changes the types of cases to which the study can generalize, so it is important to verify that findings from propensity-adjusted studies are generalizable.

**[END NL]**

One may wonder if all these manipulations of who seeks treatment lead to “fake” science. The three steps outlined earlier change observed data and essentially simulate what impact the treatment program would have had if the patients had participated equally. This equality of treatment, of course, is not the truth. Yet, the simulation is necessary to remove the bias introduced by patients’ preferences for treatment. In the end, propensity scoring allows us to simulate how things would have been if treated and untreated patients did not differ in their characteristics. These manipulations of observed data are not arbitrary or “fake” science. They are part of a systematic transformation of data that guarantees less bias.

## [H1] Balancing Through Propensity Scores

Because treatment is usually a binary variable, the propensity score is often estimated using a logistic regression model. Participation in treatment is regressed on various covariates. The accuracy of the logistic regression in removing the confounding in the data can be tested. The propensity score, whether used to match, weight, or stratify the data, should make the rate of occurrence of the covariates among treated and untreated patients the same. It is important to verify that covariates are, in fact, balanced after the use of propensity scores. For example, the analyst could verify that in strata of subjects with the same propensity score, the distribution of measured covariates will be the same in treated and untreated patients. Or, he could verify that after matching on propensity scores, treated and untreated groups have the same rate of measured covariates.

After weighting the patients using the inverse propensity scoring, treated and untreated groups should have the same rate of covariates. Typically, a statistician starts with a logistic regression, with a linear model of all covariates predicting participation in treatment. If a test indicates that some variables remain out of balance, interaction terms with the variable are introduced and the process repeated until treated or untreated groups do not differ in any covariates. These tests can, and should, be done as linear models of propensity scores may fail to balance the data. Austin (2011, 413) provides a succinct description of the iterative process involved:

**[INSERT BLOCK QUOTE]**

One begins by specifying an initial propensity score model. The comparability of treated and untreated subjects in the resultant matched sample is then assessed. If important residual systematic differences between treated and untreated subjects are found to remain, the initial propensity score model can be modified. One can modify the propensity score by including additional covariates, by adding interactions between covariates that are already in the model, or by modeling the relationship between continuous covariates and treatment status using nonlinear terms (e.g., using cubic smoothing splines). One proceeds in an iterative fashion until systematic differences in observed baseline covariates between treated and untreated subjects have either been eliminated or reduced to an acceptable level.

**[END BLOCK QUOTE]**

It is important that at each step of the iterative process, the analyst is not guided by the statistical significance of the estimated regression coefficients in the propensity score model (assuming she is using a logistic regression model). Rather, she is working toward the objective of creating a matched sample in which the distribution of observed baseline covariates is similar between treated and untreated subjects.

How to test the comparability of treated and untreated subjects in a propensity score–matched sample has been discussed extensively in the literature (Austin 2009). The simplest method is to compare the standardized difference of covariates in treated and untreated groups. The method to do so was covered in chapter 4, which focused on the comparison of means (for continuous covariates) and the comparison of rates (for discrete covariates). For continuous covariate *X*, the test of the comparison of means can be carried out using the equation  
**[INSERT EQUATION]**

.

**[END EQUATION]**

In this equation, and are the mean and sample variance of the covariate for treated patients; and are the same for the untreated group.

For a discrete variable, the estimated probability of the covariate can be calculated and used to test the assumption of balance in the treated and untreated groups, as in the equation

**[INSERT EQUATION]**

.**[END EQUATION]**

In this equation, and are the rate of occurrence of the covariate treated and untreated patients.

The standardized difference provides one of many methods for checking that the main effects of the covariates have been balanced. Careful analysis is necessary to detect any imbalances in the data (e.g., plotting the distributions, quantile-quantile plots, empirical nonparametric density plots). In addition, because propensity scoring reduces the number of patients in the analysis, it may not be reasonable to rely on tests of statistical significance. A test of statistical signiﬁcance makes sense if the number in treated and untreated groups remains high. In addition, when the main effects of the covariates are balanced, there is no guarantee that the interaction effects are balanced. It is possible that age and gender are balanced, but the interaction of gender and age is not balanced. A test of balance must be carried out for all main effects and all interaction effects. The sheer number of covariates may make the test of balance of interactions impractical. In chapter 16, we show the stratified covariate balancing method, in which all interaction terms are guaranteed to be balanced.

## [H1] Propensity Score Quintile Matching

In the matching method, the objective is to select treated and untreated patients who have similar propensity scores. There are different ways to define a match. The statistician could match a treated patient to the untreated patient with the closest propensity score. Another method is to consider treated and untreated patients to match if they are within a small span (e.g., one-quarter of the logit standard deviation of propensity score) of each other. As we match treated and untreated groups, it is important to do so without replacement—that is, the same untreated case cannot be matched to two different treated patients. Matching with replacement alters the variance of the covariate in the treated and untreated groups. Matching can be done coarsely (e.g., within five years of the patient’s age) or finely (e.g., within six months of the patient’s age). The general advice is to start with coarse matching and report the sensitivity of the findings to refinements in the match criteria. At the end of the matching, the analyst must face a crucial question: how many patients are left unmatched? In almost all cases, not all patients can be matched. If enough patients are not matched, then it may not be possible to generalize the study findings.

Once treated and untreated patients have been matched, the effect of treatment on outcomes can be estimated using a paired *t*-test. If the outcome is continuous (e.g., cost), the effect of treatment can be estimated by comparing the means of treated and untreated matched groups. For dichotomous outcomes (e.g., satisfaction with care), the effect of treatment can be estimated by comparing the rates of outcomes in treated and untreated groups. It is generally recommended that paired *t*-tests be used to examine outcomes in matched treated and untreated patients.

Nearest-neighbor matching within a specified caliper distance (so called because it measures the maximum amount that matched objects can be apart) is similar to nearest-neighbor matching, but with the further restriction that the absolute difference in the propensity scores of matched subjects must be below some prespecified threshold. Thus, for a given treated subject, one would identify all the untreated subjects whose propensity score lay within a specified distance of that of the treated subject. From this restricted set of untreated subjects, the untreated subject whose propensity score was closest to that of the treated subject would be selected for matching. If no untreated subjects had propensity scores that lay within the specified caliper distance of the propensity score of the treated subject, the treated subject would not be matched with any untreated subject. The unmatched treated subject would then be excluded from the resultant matched sample.

A variety of statistical packages exist to proceed with propensity score matching. In R, the packages MatchIt and Optmatch allow one to implement a variety of different matching methods. In Stata, the PSMATCH2 module can be used for propensity score matching. The call for propensity score matching in the MatchIt package is provided in the following box for the MFH data.

|  |
| --- |
| One easy way to test the balance among the covariates is to calculate propensity scores and stratify the data by these scores, thus making sure that residents in MFHs and contract nursing homes (CNHs) have the same propensity of joining an MFH. The results of a comparison of MFH to the weighted regression of MFH participation on various covariates can be seen in the following text. Our treatment variable is participation in an MFH, and our outcome variable is Cost. We begin by regressing participation in MFH on the other variables, excluding the outcome variable Cost.  **[LIST FORMAT]**    **[END LIST]**  We then use the model generated to predict the propensity of participation in the MFH program using the predict() function:  **[LIST FORMAT]**    **[END LIST]**  Next, we split up the propensities predicted into quintiles (five groups) using the quantile function as follows:  **[LIST FORMAT]**    **[END LIST]** |

Exhibit 13.1 shows the results when we match the MFH and CNH residents by the quintiles of propensities. For each quintile, the table reports the number of residents who were in MFHs and CNHs. Of the 16,226 residents, 3,246 fall into each propensity quintile. By definition, the number of residents that fall in each quintile is equal. Note that more of the CNH residents fall in the first quintile than MFH residents. The situation is reversed by the fifth quintile, where more MFH patients fall into the fifth quintile than CNH patients. Clearly, comparing these two programs across all quintiles is like comparing apples to oranges. Residents of the two programs are self-selecting. It makes sense to compare the two programs in each quintile. For instance, comparing the 45 residents in quintile 1 of the MFH program with the 3,201 residents in quintile 1 of the CNHs makes sense. These two programs are both serving residents who fall in the same propensity quintile—therefore, these residents do not differ on their propensity to join the programs. Thus, observed outcome differences are not confounded with differences in patients who elect to take advantage of these programs.

**[INSERT EXHIBIT]**

**Exhibit 13.1** Number of Patients Receiving Treatment Matched on Propensity Quintiles

|  |  |  |  |
| --- | --- | --- | --- |
| **Propensity Quintiles Based on Resident Characteristics** | **MFHs (Treated Group)** | **CNHs (Untreated Group)** | **Total Number of Unique Residents** |
| 1 | 45 | 3,201 | 3,246 |
| 2 | 89 | 3,156 | 3,245 |
| 3 | 168 | 3,077 | 3,245 |
| 4 | 514 | 2,731 | 3,245 |
| 5 | 1,775 | 1,470 | 3,245 |
| **Totals** | 2,591 | 13,635 | 16,226 |

**[END EXHIBIT]**

|  |
| --- |
| Exhibit 13.2 summarizes the differences between the two programs in each quintile of the propensity scores. To create the data in table 2, we used the following code. First, we created an empty table.  **[LIST FORMAT]**    **[END LIST]**  Next, we passed through the counts matrix titled “t” to extract the relevant information:  **[LIST FORMAT]**    **[END LIST]** |

**[INSERT EXHIBIT]**

**Exhibit 13.2** Cost per Follow-Up Day for MFH and CNH

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Propensity Quintile** | **Medical Foster Home ($)** | **Nursing Home ($)** | **Difference ($)** | **Number of Patients** |
| 1 | 2.71 | 87.98 | −85.27 | 3,246 |
| 2 | 4.44 | 77.80 | −73.36 | 3,245 |
| 3 | 11.31 | 78.58 | −67.26 | 3,245 |
| 4 | 31.82 | 72.22 | −40.40 | 3,245 |
| 5 | 109.62 | 40.54 | 69.08 | 3,245 |

**[END EXHIBIT]**

Exhibit 13.2 shows the difference in cost in various propensity quintiles. These differences range from a low of −$40.40 in quintile 4 to a high of $69.08 in quintile 5. These cost differences do not result from residents’ propensity to select MFHs over CNHs. We controlled for these differences by calculating the cost differences in each quintile. On the whole, it seems that the MFH saves money, but exceptions exist. For residents who fall in quintile 5, it costs more. For all others, it costs less.

The use of quintiles to match propensity scores assumes that all relevant characteristics of the patient are taken into account in the calculation of the propensity score. This is often the case. When there is a large number of characteristics, and when these characteristics interact and the logistic regression that derived the propensity score does not include the interactions, the propensity score may not balance the combination of patient characteristics. Some differences continue to exist in the nature of residents who select MFHs or CNHs. Particular combinations of patient characteristics may remain unbalanced across the two treatments.

## [H1] Propensity Scoring Weighting

One way to use propensity scoring to remove the confounding is to use every case but to weight cases with a low propensity to participate in treatment heavily and cases commonly participating in treatment sparingly. In this fashion, covariates will be equally likely to be in treated and untreated groups. The approach is called inverse probability of treatment weighting (IPTW). Given that many software packages allow weighted regression, the IPTW is easily implemented.

The IPTW weights depend on whether the patient is in the treated or untreated group. If Ti is 1 when patient *i* is treated and 0 otherwise, then the weights for both treated and untreated groups, , can be calculated as the following:

**[INSERT EQUATION]**

.**[END EQUATION]**

In this equation, is the propensity of participating in treatment for case *i*. Note that for treated patients, The equation for the weights simplifies to

**[INSERT EQUATION]**

**[END EQUATION]**

In short, for treated patients, the weights are the inverse of the propensity of participating in treatment. For patients who did not receive treatment (i.e., when the weight is the inverse of probability of not being in treatment, written as   
**[INSERT EQUATION]**

.**[END EQUATION]**

Thus, the weight equation behaves differently for treated and untreated patients, but in principle, it increases the rate of participation of patients who are unlikely to participate in treatment and reduces the rate of common participants. These adjustments improve the chances that treated and untreated patients have the same rates of different covariates.

Once the weights have been set, weighted regression can be used to estimate the treatment effect. Weighted regression of the outcome on treatment and covariates has the added advantage of removing residual differences that might have remained despite propensity scoring. Alternatively, the treatment effect can also be estimated using the following formula (Lunceford and Davidian 2004):  
**[INSERT EQUATION]**

.**[END EQUATION]**

In this equation, n denotes the number of subjects, and outcome in the ith patient.

If we use the data in exhibit 13.1, residents who fall in quintile 1 have chance of receiving MFH care. Residents in quintile 1 have chance of receiving CNH services. The weight for the 45 residents in quintile 1 of the MFH program is 1 ÷ 0.014, which counts the 45 cases as 3,246 weighted cases. The weight for 3,246 quintile 1 residents of the CNH is 1 ÷ 0.986, which also results in 3,246 weighted cases. Once weighted by the inverse propensity score, the number of residents in quintile 1 is the same across the two programs. Now, we can focus on outcomes for these weighted cases without worrying that an unequal number falls in quintiles.

## [H1] Double Regression

Analysis with propensity scoring involves two regressions. The first step uses logistic regression to predict the propensity of being in treatment from various covariates. This propensity score is used to calculate inverse propensity treatment weights. These weights are then used in the second regression: outcomes of treatment are examined in a weighted regression of outcome on treatment plus any covariates or combination of covariates that have not balanced.

## [H1] Example for Weighted Propensity Scoring

We demonstrate these procedures by examining the effectiveness of antidepressants, in particular citalopram. In the data available through experiments conducted with funding from the National Institute of Mental Health, patients were given Buspirone, and its impact on the remission of depression symptoms was reported. These data also report patients’ medical history at baseline. It is possible that patients’ history affects their selection of citalopram as their antidepressant. To examine whether that is the case, we run a regression of Buspirone on various baseline predictors.

|  |
| --- |
| The R code for conducting the regression is as follows:  **[LIST FORMAT]**  glm(formula = Buspirone ~ Gender + RiskOfSuicide + Heart + Vascular +  Haematopoietic + Eyes\_Ears\_Nose\_Throat\_Larynx + Gastrointestinal +  Renal + Genitourinary + Musculoskeletal\_Integument + Neurological +  Psychiatric\_Illness + Respiratory + Liver + Endocrine + Alcohol +  Amphetamine + Cannabis + Opioid + Panic + Specific\_Phobia +  Social\_Phobia + OCD + PTSD + Anxiety + Borderline\_Personality +  Dependent\_Personality + Antisocial\_Personality + Paranoid\_Personality +  Personality\_Disorder + Anorexia + Bulimia + Cocaine, family = "binomial",  data = mydata)  **[END LIST]**  The code instructs the computer to regress taking Buspirone on dozens of factors. The code also specifies that the binomial distribution should be used, which leads to a logistic regression. The code also says where to find the data. |

The result of the logistic regression is provided in exhibit 13.3. It shows that patients who took Buspirone were more likely to be male; to have eye, ear, nose, throat, and larynx (*p* = 0.01) conditions, neurological (*p* = 0.03) problems, and respiratory (*p* = 0.00) conditions; and to have used amphetamines (*p* = 0.02). They were less likely to have heart (p = 0.05) or musculoskeletal integument (*p* = 0.01) conditions and to abuse alcohol (*p* = 0.02). Clearly, the assignment to take Buspirone was not random. Many baseline differences are confounded with the impact of Buspirone on remission of depression symptoms. These baseline differences necessitate the use of propensity scores to remove the confounding.

**Exhibit 13.3** Propensity to Take Buspirone

|  |
| --- |
| **[LIST FORMAT]**  Deviance Residuals:  Min 1Q Median 3Q Max  -0.5723 -0.3286 -0.2985 -0.2699 2.9474  Coefficients:  Estimate Std. Error z-value Pr(>|z|)  (Intercept) -2.98993 0.07157 -41.775 < 2e-16 \*\*\*  Gender2 -0.25602 0.06990 -3.663 0.00025 \*\*\*  RiskOfSuicide -0.73781 0.41598 -1.774 0.07612 .  Heart -0.19296 0.09709 -1.987 0.04687 \*  Vascular 0.11197 0.08031 1.394 0.16325  Hematopoietic 0.04184 0.11610 0.360 0.71858  Eyes\_Ears\_Nose\_Throat\_Larynx 0.17215 0.06925 2.486 0.01292 \*  Gastrointestinal -0.04829 0.07069 -0.683 0.49448  Renal -0.02022 0.13873 -0.146 0.88412  Genitourinary -0.14076 0.08451 -1.666 0.09580 .  Musculoskeletal\_Integument -0.19963 0.07127 -2.801 0.00510 \*\*  Neurological 0.15715 0.07458 2.107 0.03512 \*  Psychiatric\_Illness 0.02985 0.10419 0.287 0.77449  Respiratory 0.29625 0.07004 4.230 2.34e-05 \*\*\*  Liver 0.02673 0.11493 0.233 0.81606  Endocrine 0.12087 0.07898 1.530 0.12592  Alcohol -0.41872 0.18291 -2.289 0.02206 \*  Amphetamine 0.91232 0.40340 2.262 0.02372 \*  Cannabis -0.41167 0.38906 -1.058 0.29000  Opioid -13.32932 326.57719 -0.041 0.96744  Panic -0.10713 0.17758 -0.603 0.54634  Specific\_Phobia -0.03045 0.44841 -0.068 0.94586  Social\_Phobia 0.22176 0.17367 1.277 0.20165  OCD 0.45623 0.36371 1.254 0.20970  PTSD 0.15575 0.14179 1.098 0.27201  Anxiety -0.31096 0.15951 -1.949 0.05124 .  Borderline\_Personality 0.31052 0.40077 0.775 0.43845  Dependent\_Personality -14.29934 446.52547 -0.032 0.97445  Antisocial\_Personality -14.47118 593.82834 -0.024 0.98056  Paranoid\_Personality -14.57253 1604.64909 -0.009 0.99275  Personality\_Disorder -14.41464 467.78133 -0.031 0.97542  Anorexia -14.43478 1317.92657 -0.011 0.99126  Bulimia -14.42094 549.75084 -0.026 0.97907  Cocaine -12.30392 180.83433 -0.068 0.94575  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 8222.7 on 22212 degrees of freedom  Residual deviance: 8089.8 on 22179 degrees of freedom  AIC: 8157.8  Number of Fisher Scoring iterations: 16  **[END LIST]** |

We use inverse propensity weights to remove the confounding in the data. Doing so has three steps. First, the analyst must attach the propensity scores to the data frame, so that the predicted propensity scores are available. Next, the propensity weights are calculated. Last, the weights are used to balance the data and remove the confounding. The following R code shows the step.

|  |
| --- |
| In the first step, the R code attaches propensity scores to the data frame:  **[LIST FORMAT]**  mydata$psvalue<-predict(propscore,type="response")  **[END LIST]**  In the second step, the weights are calculated. The “If Else” command is used to assign separate weights for patients who took Buspirone (treated cases) and those who did not (untreated controls):  **[LIST FORMAT]**  mydata$weight<-ifelse(mydata$Remission==1,1/mydata$psvalue,1/(1-mydata$psvalue))  **[END LIST]**  In the last step, a weighted regression of remission rates on various baseline predictors and participation in treatment is carried out. In contrast to our calculation of propensity scores, the dependent variable is now remission rates, not use of Buspirone. After the tilde, the R code lists the independent variables. Buspirone is now listed as an independent variable. Baseline variables are also listed. The last part of the R code tells the computer to use the field called “Weight” in the data frame for the weights in the regression:  **[LIST FORMAT]**  lm(formula = Remission ~ Buspirone + Gender + RiskOfSuicide +  Heart + Vascular + Hematopoietic + Eyes\_Ears\_Nose\_Throat\_Larynx +  Gastrointestinal + Renal + Genitourinary + Musculoskeletal\_Integument +  Neurological + Psychiatric\_Illness + Respiratory + Liver +  Endocrine + Alcohol + Amphetamine + Cannabis + Opioid + Panic +  Specific\_Phobia + Social\_Phobia + OCD + PTSD + Anxiety +  Borderline\_Personality + Dependent\_Personality + Antisocial\_Personality +  Paranoid\_Personality + Personality\_Disorder + Anorexia +  Bulimia + Cocaine, data = mydata, weights = weight)  **[END LIST]** |

The regression results are provided in exhibit 13.4. There is a statistically significant relationship between use of Buspirone and reduction in depression symptoms (Coefficient = −0.092, *p*< 0.000). This estimate does not reflect the main effects of various baseline parameters, as these were statistically controlled through the propensity scores. Also note that several variables, including variables that were not significant predictors of propensity to take Buspirone, are significant predictors of remission. Patients who have endocrine conditions, alcohol abuse, opioid use, specific phobias, obsessive-compulsive disorder, bulimia, cocaine use, or various personality disorders are less likely to have symptom remission, independent of whether they take Buspirone. The unconfounded impact of Buspirone can be used to compare it to other antidepressants and decide which antidepressant is best for a patient.

**[INSERT EXHIBIT]**

**Exhibit 13.4** Weighted Regression of Remission on Buspirone and Other Baseline Predictors

|  |
| --- |
| **[LIST FORMAT]**  Weighted Residuals:  Min 1Q Median 3Q Max  -1.08488 -1.02510 -1.01905 -0.00015 1.63261  Coefficients:  Estimate Std. Error t-value Pr(>|t|)  (Intercept) 9.998e-01 9.369e-05 10670.851 < 2e-16 \*\*\*  Buspirone -9.161e-02 9.997e-03 -9.164 < 2e-16 \*\*\*  Gender2 -2.624e-07 3.681e-05 -0.007 0.994312  RiskOfSuicide 4.695e-04 2.404e-04 1.953 0.050870 .  Heart -5.019e-05 4.746e-05 -1.058 0.290266  Vascular -2.440e-05 7.122e-05 -0.343 0.731891  Hematopoietic 1.473e-04 1.282e-04 1.149 0.250574  Eyes\_Ears\_Nose\_Throat\_Larynx -8.665e-05 5.998e-05 -1.445 0.148570  Gastrointestinal -1.004e-04 5.612e-05 -1.789 0.073626 .  Renal 1.487e-04 1.060e-04 1.403 0.160606  Genitourinary 2.920e-05 5.971e-05 0.489 0.624902  Musculoskeletal\_Integument -8.524e-05 5.924e-05 -1.439 0.150196  Neurological 6.855e-05 3.846e-05 1.782 0.074714 .  Psychiatric\_Illness -1.658e-05 3.453e-05 -0.480 0.631175  Respiratory 4.869e-05 4.829e-05 1.008 0.313369  Liver -7.085e-05 3.823e-05 -1.853 0.063891 .  Endocrine -1.390e-04 6.303e-05 -2.205 0.027461 \*  Alcohol -5.170e-05 4.220e-05 -1.225 0.220522  Amphetamine -8.204e-05 6.456e-05 -1.271 0.203828  Cannabis -7.096e-05 5.418e-05 -1.310 0.190352  Opioid 3.789e-04 1.577e-04 2.403 0.016285 \*  Panic -5.625e-05 1.104e-04 -0.510 0.610341  Specific\_Phobia -1.131e-01 3.152e-02 -3.588 0.000333 \*\*\*  Social\_Phobia -1.017e-05 1.149e-04 -0.088 0.929518  OCD -3.651e-04 1.663e-04 -2.196 0.028089 \*  PTSD 1.799e-04 9.275e-05 1.940 0.052416 .  Anxiety -1.224e-04 6.743e-05 -1.815 0.069491 .  Borderline\_Personality -3.398e-01 4.987e-02 -6.813 9.81e-12 \*\*\*  Dependent\_Personality 3.595e-04 1.580e-04 2.275 0.022893 \*  Antisocial\_Personality 4.228e-04 1.894e-04 2.232 0.025623 \*  Paranoid\_Personality -9.240e-01 3.610e-01 -2.559 0.010494 \*  Personality\_Disorder 2.629e-04 1.123e-04 2.340 0.019270 \*  Anorexia 2.455e-04 1.403e-04 1.750 0.080191 .  Bulimia 3.138e-04 1.240e-04 2.531 0.011391 \*  Cocaine 2.098e-04 8.149e-05 2.575 0.010034 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.8828 on 22178 degrees of freedom  Multiple R-squared: 0.007241, Adjusted R-squared: 0.005719  F-statistic: 4.758 on 34 and 22178 DF, p-value: < 2.2e-16  **[END LIST]** |

## [END EXHIBIT]

## [H1] Verification of Propensity Scores

It is important to confirm that inverse propensity weights have removed the effects of various predictors. This is typically done by recalculating the propensity score in a weighted regression. If the weights have adequately removed the effects of baseline values, none of these variables should be a statistically significant predictor of treatment.

## [H1] Overlap and Related Concepts

Confounding can be removed through matching (not shown in this chapter but described in chapter 15 on case/control matching), by division of propensity scores into quintiles, and by inverse propensity weights. In propensity scoring, one is always concerned with the adequacy of matches or with the use of extreme weights. *Adequate matching* refers to overlap between cases and controls. Typically, there are many controls (the untreated group), and the task is to match as many cases (from the treated group) to controls. The percentage of cases that are matched is called *overlap between cases and controls*. When the overlap is low, the analyst may become concerned that the findings will not generalize to the unmatched cases. If too many such cases exist, the entire process is questionable and the findings are of little use.

Similarly, if we are trying to use weights, there is concern with extreme weights that count one case too many times. If patients differ radically in their treatment selections, there may be one treated patient for many untreated patients; in this case, it is possible that the weights will be extreme values. The outcome for these treated patients will have a large influence on regression findings. If the weights are inaccurate, if the sample was somehow unusual, an error in estimating the outcome in these highly weighted cases may lead to large errors overall. It is important to check the weights to ensure that no single case is given overwhelming importance.

## [H1] Summary

This chapter introduced the use of logistic regression in propensity scoring. We showed different ways to incorporate propensity scoring into the analysis. As a consequence of propensity scoring, statisticians can calculate the unconfounded estimate of treatment.

## [H1] Supplemental Resources

A problem set, solutions to problems, multimedia presentations, SQL code, and other related material are on the course website.

**[H1] References**

Austin, P. C. 2011. “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.” *Multivariate Behavioral Research* 46 (3): 399–424.

———. 2009. “The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies.” *Medical Decision Making* 29 (6): 661–77.

Grabowski, D. C., D. G. Stevenson, D. J. Caudry, A. J. O'Malley, L. H. Green, J. A. Doherty, and R. G. Frank. 2017. “The Impact of Nursing Home Pay‐for‐Performance on Quality and Medicare Spending: Results from the Nursing Home Value‐Based Purchasing Demonstration.” *Health Services Research* 52 (4): 1387–408.

Kim, C. W., S. H. Baik, Y. H. Roh, J. Kang, H. Hur, B. S. Min, K. Y. Lee, and N. K. Kim. 2015. “Cost-Effectiveness of Robotic Surgery for Rectal Cancer Focusing on Short-Term Outcomes: A Propensity Score-Matching Analysis.” *Medicine* 94 (22): e823.

Lunceford, J. K., and M. Davidian. 2004. “Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study.” *Statistics in Medicine* 23 (19): 2937–60.

Marsteller, J. A., Y. J. Hsu, C. Gill, Z. Kiptanui, O. A. Fakeye, L. D. Engineer, D. Perlmutter, N. Khanna, G. B. Rattinger, D. Nichols, and I. Harris. 2018. “Maryland Multipayor Patient-Centered Medical Home Program: A 4-Year Quasiexperimental Evaluation of Quality, Utilization, Patient Satisfaction, and Provider Perceptions.” *Medical Care* 56 (4): 308–20.

Scappaticci, G. B., B. L. Marini, V. R. Nachar, J. R. Uebel, V. Vulaj, A. Crouch, D. L. Bixby, M. Talpaz, and A. J. Perissinotti. 2018. “Outcomes of Previously Untreated Elderly Patients with AML: A Propensity Score-Matched Comparison of Clofarabine vs. FLAG.” *Annals of Hematology* 97 (4): 573–84.