**Chapter 16**

**Stratified Covariate Balancing**

# [H1] Learning Objectives

* Identify weights that would balance the main effect and interaction among covariates
* Calculate the weighted, unconfounded impact of treatment on outcome
* Remove confounding in electronic health records without access to statistical software
* Calculate the common odds ratio across strata
* Calculate overlap between cases and controls for exact and partial matches
* Improve overlap among cases and controls through adding synthetic controls
* Improve overlap among cases and controls through use of expected values

# [H1] Key Concepts

* Stratified covariate balancing
* Percent overlap
* Partial stratification
* Data balancing
* Common odds ratio
* Weighted impact
* Parents in Markov Blanket
* LASSO (least absolute shrinkage and selection operator) regression
* Switching distributions

# [H1] Chapter at a Glance

In the chapter on propensity scoring, we acted as if the assessment of propensity scores is a relatively easy task. We suggested that all a statistician needs to do is regress treatment variables on the covariates. We lied. The truth is that the procedure is more complex—such regressions must exclude covariates that occur after treatment (a relatively easy task) and include interactions among covariates (a much more difficult task). Specifying interactions is often difficult, if not impossible, because many combinations of variables could happen. Many analysts balance the data using main effects, which could be misleading. This chapter provides a simpler method that automatically balances the interactions among the covariates. Thus, it overcomes one of the most difficult part of data balancing. The approach is called stratified covariate balancing (see Alemi, ElRafey, and Avramovic 2016). It provides a set of weights that can replace the inverse propensity weights. It makes data balancing relatively easy because the weights are assessed from formulas without use of regression. It provides a procedure that can be done within electronic health records (EHRs), where statistical software are typically not available.

# [H1] Introduction

Analysts commonly need to estimate the impact of a variable (e.g., treatment) on another (e.g., an outcome) while statistically controlling for multiple covariates (e.g., comorbidities). The impact of treatment is confounded with a host of other variables, including a patient’s medical history, past medication, current medication, allergies, and even genes. Statistical procedures can be used to separate out the effect of treatment from other comorbidities.

Statisticians typically address confounding through randomization. In 1983, Paul R. Rosenbaum and Donald B. Rubin proposed methods for removing confounding in observational data. They proposed the use of propensity scoring to balance rates of occurrence of covariates among treated and untreated subjects. Since then, different methods of propensity scoring have been proposed, including methods for matching, subclassification, weighting, regression, likelihood, or combinations of approaches. Chapter 13 showed how propensity scores can be calculated from the regression of treatment on covariates. It also showed how these scores can be used to balance data. This chapter shows an alternative approach to balancing data: stratified covariate balancing.

# [H1] The History of Stratification

Stratification procedures have been available since the 1950s but fell out of popular use because of difficulty with stratifying high-dimensional data (data with many features of variables—sometimes in the thousands). In recent years, progress with feature reduction, as well as network modeling, have eased the problems. Chapter 16 introduces the concept of stratified covariate balancing, and later chapters will show how this method can be used in network modeling and causal analysis. Stratified covariate balancing divides the data into strata and balances the data in each stratum.

The procedure for finding the strata is not statistical but analytical. It does not involve parameter estimation. It does not assume distributions. Essentially, it is a divide and conquer approach, in which data are divided into subgroups, and treated and untreated patients are contrasted within the subgroups. The analyst begins by searching the data for all combinations of covariates. Each unique combination is considered one stratum or subgroup. In each stratum, the levels of covariates are fixed. Treated and untreated patients within the subgroups share the same strata; therefore, the differences within the strata cannot result from covariates. Treated patients are referred to as *cases*, and untreated patients are referred to as *controls*. The comparison of cases to controls in the same stratum estimates the impact of treatment while holding values of the covariates at the same level. In effect, the procedure allows the analyst to hold covariates constant and examine the effect of changes in treatment variable.

# [H1] Combination of Covariates

In most databases, the search for a combination of covariates is practical. In theory, the possible combination of *k* binary covariates is 2k, which, depending on the size of *k*, may exceed the computer’s capability. However, most combinations do not occur in the data. The observed set of combinations is many times fewer than the theoretically possible set of combinations. Even in multidimensional, massive data, the number of possible combinations is relatively small. For example, Levy and colleagues (2015) analyzed 1.3 million records for 11 variables. Theoretically, there should be 211 = 2,048 possible strata, but they observed only 418 unique combinations of 11 variables. Nearly 79 percent of the possible combinations of covariates never occurred in the data, despite its large scale.

There are many reasons for this outcome. Certain combinations are not possible (e.g., pregnant male). Other combinations are possible but unlikely (e.g., nursing home resident who is unable to sit but can use the toilet). No matter why certain combinations do not occur, the lack of occurrence simplifies the analysis. The number of strata that need to be examined is radically reduced, making the analysis more practical.

# [H1] Impact of Treatment on Binary Outcomes

Stratification is essentially subgroup analysis; it divides the data into smaller sets. The idea is to separate the data into several subgroups, in which the effect of confounding variables is constant and the impact of treatment can be assessed. These subgroups are called *strata*, and the process of dividing the data into the subgroups is called *stratification*. In general, the data will be stratified into *k* subgroups, each holding a unique combination of confounding variables constant.

In exhibit 16.1, the data are organized as cases and controls, with cases showing treated patients and controls showing untreated patients. A stratum is composed of *n* covariates shared in both cases and controls. For example, the covariates could be age, gender, and the comorbidities of the patient. Cases could be patients who received an antidepressant. The outcome could be remission of depression symptoms. Each stratum reports the impact of cases and controls on an outcome, shown as *y*.

In exhibit 16.1, the cell entries are a count of distinct patients. In each, contrasting cases to controls allows us to estimate the impact of treatment on the outcome, without concern for the covariates. Thus entry *ai* shows the number of cases with the desired outcome and the same characteristics as controls. For example, *ai* shows the number of patients who took the antidepressant, had remission of their depression symptoms, and had the same comorbidities as the control patients who did not take the antidepressant.

**Exhibit 16.1** Outcomes in ith Stratum, *i* = 1, . . . , *k*

|  |  |  |  |
| --- | --- | --- | --- |
| **Patients’ Characteristics** | | **Outcome *y* = 1** | **Outcome *y* = 0** |
| Same n Covariates for Cases and Controls | Cases  (*x* = 1) | *ai* | *bi* |
| Controls (*x* = 0) | *ci* | *di* |

To estimate the impact of treatment, we follow procedures for the analysis of stratified case-control design. The chi-square test for homogeneity is used to see whether, across strata, a common odds ratio exists. If cases and controls are counted as they are in exhibit 16.1, the chi-square test of homogeneity of treatment impact across strata is calculated as

In this equation, , , and . The counts, , , and are defined in exhibit 16.1. The symbol is referred to as the *squiggly* or *tilde* and is used to indicate distribution of the statistics. The term says that the calculated value has a chi-square distribution with s− 1 degrees of freedom, where *s* is the number of strata.

If the hypothesis that the odds ratios are the same across the strata is rejected, the common odds ratio does not exist. In this case, more than one common odds ratio needs to be estimated for different part of the strata. Otherwise, a common odds ratio does exist, and can be calculated as

In this equation, , , , and are measured as shown in exhibit 16.1, and is the total number of cases in the strata.

# [H2] Application to Employees’ Contribution to Satisfaction Ratings

An example can demonstrate the calculations. Exhibit 16.2 shows the satisfaction of patients with teams of providers. The purpose of analysis is to examine the satisfaction associated with each member of the team. Exhibit 16.2 shows the team’s data. This data can be arranged in two ways, depending on whether we are interested in the impact of the physicians or the nurses. In exhibit 16.2, the table on the left shows two strata, in which the impact of nurses on complaints can be studied. In the table on the right of exhibit 16.2, the nurses are considered covariates, and physician’s impact on complaints can be studied.

**Exhibit 16.2** Stratification of Satisfaction with Teams of Providers

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Impact of Nurses on Complaints** | | | | |  | **Impact of Physicians on Complaints** | | | | |
| **Strata** | **Physicians as Covariates** | **Nurses as Treatment** | **Complaints as Outcome** | |  | **Strata** | **Nurses as Covariates** | **Physicians as Treatment** | **Complaints as Outcome** | |
| **Yes** | **No** |  | **Yes** | **No** |
| 1 | Doctor A | Nurse C | 53 | 424 |  | 1 | Nurse C | Physician A | 53 | 424 |
| Nurse D | 11 | 37 |  | Physician B | 0 | 16 |
| 2 | Doctor B | Nurse C | 0 | 16 |  | 2 | Nurse D | Physician A | 11 | 37 |
| Nurse D | 4 | 139 |  | Physician B | 4 | 139 |

We will begin with the left-hand table. We start the analysis by checking to see where the odds ratios in different strata have the same value. To calculate the statistic for the test of homogeneity, we need to first calculate the log of the odds ratio for each of the two strata in the following manner: ,

Note that in stratum 2, we had to add 0.5 to each cell in the strata because the log of zero is not defined,. Next we calculate

, and

The average log of odds ratios is calculated as

Given these component calculations, we can now calculate the statistic for testing the homogeneity of the odds ratio across the strata:

.

The statistic is calculated to be 0.5. It has a chi-square distribution. Since it was calculated from two strata, then it has 2 minus 1 degrees of freedom. The hypothesis that odds ratios in these strata are the same is not rejected. This is good news. It tells us that the common odds ratio may exist. Using the formula presented earlier, the common odds ratio can now be calculated as

This common odds ratio tells us that changing from nurse C to nurse D changes the number of complaints by 0.40 times, independent of physician partners. In essence, nurses contribute to reducing the number of complaints.

# [H1] Impact of Treatment on Continuous Outcomes: Difference Models

The common odds ratio can be calculated only when the outcome is a binary variable. When the outcome is continuous, two different approaches can be used to assess the effect of treatment. In one approach, the difference within strata are calculated and then aggregated to assess the average impact. The second approach weighs the data so that covariates are balanced in treated and untreated groups. This section describes the difference model, and the next section discusses the use of weights.

For each stratum, the difference is calculated as

In this equation, is a set of weights proportional to the frequency of cases in the strata: These weights not used to balance the data but to give more influence to larger strata. The difference in each stratum is shown as and calculated as the average of treated patients across the strata minus average of controls across the same strata, or

The average of cases, or treated patients, is calculated as . The average of untreated patients, or the controls, is calculated as . In the strata, the covariates are the same for cases and controls. Therefore, the difference of the two averages reports the impact of treatment on outcome after controlling for the covariates.

The average treatment effect, , has a Student’s *t-*distribution with s ˗ 1 degrees of freedom, and the *t*-statistic is calculated as . The standard deviation of the treatment effect is calculated as .

# [H1] Impact of Treatment on Continuous Outcomes: Weighted Data

If the analysis has to be carried out at the patient level and not within strata *i*, a weighting procedure can produce the same result. To get a sense of the logic of why weighting makes sense, it is helpful to think of it as resampling. Each time we sample, we find a different frequency of covariates, sometimes higher and other times lower. Weighting also changes the frequency of the covariates without forcing the analyst to go through the physical process of resampling. Imagine resampling data until, if we are very lucky, we hit on a sample in which the covariates do not differ among treated and untreated groups. In this implausibly fortunate scenario, covariates occur at the same rate among cases and controls. These resampled data can be used to evaluate the impact of treatment with no consideration of confounding.

Of course, we are never so lucky as to balance the data through physically carrying out resampling. We can sample millions of times and not run into a sample in which all the covariates are balanced. Alternatively, if we think of resampling as weighting, then we can select weights that accomplish the needed balance, thus removing the need for luck. The weights will balance the data and show us the rare situation in which resampling would have accomplished the same pattern.

In resampled data, the new count of cases and controls is shown as a weighted product of the original sample counts. A set of weights that guarantees that covariates in cases and controls have the same rate is

In this formula, is 1 when treated and 0 otherwise; furthermore,*, ,* , and are measured as shown in exhibit 16.1. For the matched case, where *Ti*= 1, the weight is 1—in essence matched cases are not weighted. For the matched control, where *Ti*= 0, the weight is the ratio of number of cases divided by the number of controls. When either the case or the control does not have a match, the weights are set to 0.

Because each stratum is mutually exclusive, the weights assigned in this fashion do not contradict each other. One set of weights does not balance one covariate but creates new imbalances in another. The weights balance the unique combination of the covariates that occur in each stratum.The weights ensure that cases and controls in each stratum, and by extension the combination of covariates in the strata, have equal frequency. Because combinations of covariates are occurring at the same rate in cases and controls, any mixture of these combinations, including the subset with just one covariate, will occur with equal frequency. This means that not only the main effect of the covariate but also all interactions among the covariates occur equally frequently in treated and untreated groups.

The success of weighting procedures in balancing data has led to other steps that can further simplify the process. One approach is to replicate the effect of weights without actually estimating and using them. Weights ensure that the distribution of cases and controls are the same. This match can also be ensured by switching the distribution of controls with their corresponding cases. First, data are divided into cases and controls within different strata. Then, the number of controls within the strata is set to the number of cases within the same strata, accomplishing what would have happened if the controls were multiplied by weights. This method is referred to as *switching distributions* and accomplishes the goals of the weighting procedure without first estimating the weights. We show examples of switching distributions in chapter 17.

# [H1] Comparison to Propensity Scoring

In 2016, I and two fellow scholars, Amr ElRafey and Ivan Avramovic, simulated a number of data sets. We then examined the performance of stratified covariate balancing and propensity scoring when there were significant interactions among the covariates. In high-dimensional, massive data, it is not practical to model all interactions in the variables. At best, only pair-wise interactions are modelled. Initially, when there was not much interaction among the covariates, pair-wise propensity scoring and stratified covariate balancing performed similarly.

As we used higher-order interaction terms to generate the outcome, the stratified covariate balancing method maintained its accuracy, but pair-wise propensity scoring method had increasing error. No matter how many interaction terms were used to generate the outcome, stratified covariate balancing was able to estimate the impact of treatment relatively accurately. Propensity scoring was not able to do so. The reason for the success of stratified covariate balancing is quite simple: its weights are based on a combination of covariates. Thus, the weights are based on observed interactions in the data.

# [H2] Example: Impact of Feeding Disability on Mortality

To demonstrate the use of stratified covariate balancing, Levy and colleagues (2015) used a large data set of 1,329,260 assessments of 296,051 residents in nursing homes administered by the US Department of Veterans Affairs (VA). They wanted to examine the relationship between feeding disability and six-month mortality. The mean age of the residents was 74.36 years at time of first assessment. The majority were white (79.88 percent) and male (96.34 percent), typical of patients receiving care in the VA system. An average resident had 6.15 (standard deviation = 8.76) assessments that included information on nine disabilities. These included bathing (B), walking (W), grooming (G), dressing (D), toileting (T), bowel continence (L), transfer (S), urinary continence (U), and feeding (F) disabilities. The dependent variable was mortality within six months of the assessment.

There were 418 unique combinations (strata) of disabilities for which at least one case with feeding disability and one control without were present. Exhibit 16.3 lists the 20 most frequent strata in our data. For example, stratum 10 shows males, 40–65 years old, without any other disability. In this stratum, 4,113 had feeding disability and 11,173 did not. In each row, cases and controls occur with different frequency; the weighting procedure resets these frequencies so that the cases and controls have the same frequencies. For example, the cases in stratum 10 will have a weight of 1 and controls will have a weight of 4,113 ÷ 11,173 = 0.368. Using these two weights guarantees that strata occur equally among patients with and without feeding problems. If the strata occur equally, then any particular covariate also occurs equally, as these covariates are the sum of a subset of strata. In addition, in each stratum, the mortality rate for cases and controls can be compared, as these values are calculated for the same type of resident, having same age, same gender, and same disabilities. Thus, mortality rates are calculated independent of covariates.

**Exhibit 16.3** Top 20 Most Frequent Strata

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **k** | **Age** | **Male** | **Disabilities** | **Cases**  **Unable to Eat, *X* = 1** | | **Matched Controls**  **Able to Eat, *X* = 0** | | |
| **Total,** | **Number Dead,** | **Total,** | **Number Dead,** | **Weight, *wi0*** |
|
| 1 | 65–85 | M | SGTBWDL | 36,677 | 12,831 | 17,862 | 4,253 | 2.053 |
| 2 | 40–65 | M | SGTBWDL | 19,317 | 9,787 | 10,739 | 3,512 | 1.79 |
| 3 | 65–85 | M | SGTBWD | 14,494 | 3,118 | 7,456 | 1,153 | 1.944 |
| 4 | 85+ | M | SGTBWDL | 11,336 | 3,951 | 22,220 | 5,436 | 0.51 |
| 5 | 40–65 | M | SGTBWD | 10,987 | 3,263 | 6,318 | 1,358 | 1.739 |
| 6 | 65–85 | M | GTBWD | 6,386 | 3,275 | 3,032 | 1,121 | 2.106 |
| 7 | 65–85 | M | GTBWDL | 5,101 | 2,192 | 9,524 | 2,544 | 0.536 |
| 8 | 40–65 | M | GTBWD | 4,592 | 982 | 7,283 | 1,226 | 0.631 |
| 9 | 40–65 | M | GTBWDL | 4,465 | 3,210 | 2,002 | 1,017 | 2.23 |
| 10 | 40–65 | M |  | 4,113 | 762 | 11,173 | 1,695 | 0.368 |
| 11 | 85+ | M | GTBWDL | 4,035 | 2,016 | 26,189 | 7,938 | 0.154 |
| 12 | 85+ | M | GTBWD | 4,027 | 2,031 | 9,406 | 3,167 | 0.428 |
| 13 | 65–85 | M | GTBD | 3,362 | 871 | 6,647 | 1,209 | 0.506 |
| 14 | 85+ | M | SGTBWD | 2,475 | 1,196 | 5,033 | 1,749 | 0.492 |
| 15 | 65–85 | M |  | 2,305 | 918 | 13,097 | 2,990 | 0.176 |
| 16 | 65–85 | M | GB | 1,744 | 607 | 9,122 | 1,897 | 0.191 |
| 17 | 40–65 | M | GB | 1,629 | 647 | 3,186 | 666 | 0.511 |
| 18 | 65–85 | M | GBW | 1,534 | 566 | 4,038 | 998 | 0.38 |
| 19 | 40–65 | M | GBW | 1,406 | 362 | 782 | 124 | 1.798 |
| 20 | 65–85 | M | B | 1,317 | 384 | 100,237 | 12,786 | 0.013 |

*Note*: M = male, B = unable to bathe, W = unable to walk, G = unable to groom, D = unable to dress, T = unable to toilet, L = bowel incontinent, S = unable to transfer, U = urinary incontinent.

It is informative to examine the rate of occurrence of the covariates after weighting. We expect that among patients with and without feeding disability, the rate of the covariates will be the same. Exhibit 16.4 shows the odds of various covariates co-occurring with “unable to eat.” When the sample was not weighted, these odds varied and were seldom 1 to 1. After the sample was weighted, the odds of all the co-occurring covariates were 1 to 1—they were all balanced. For example, before weighting, residents who were unable to eat were more likely to also have transfer disabilities than residents who were able to eat. The weighting procedure removed the differences in transfer disabilities. After weighting, both residents who were able to eat and those who were not had the same rate of transfer disabilities, yielding an odds ratio of 1 to 1.

Though exhibit 16.4 does not show it, other combinations of the listed covariates would also have an odds ratio of 1 to 1. The stratified covariate balancing procedure removes confounding from not only the main effects of the covariates but also from any combination of the covariates. Herein lies the real advantage of using stratified covariate balancing: the weights remove the effect of the combination of covariates without requiring us to exame interaction terms.

**Exhibit 16.4** Odds of Occurrence of the Covariate for Residents Who Are Able or Unable to Eat

Given that the data were balanced, the next step was to calculate the unconfounded odds of mortality for residents who were unable to eat. Note that in each stratum, the weighting procedure does not change the calculation of the common odds ratio, as the weights in the denominator and the numerator cancel each other out. But across strata, the estimate of the common odds ratio is very different from the unstratified estimate. The results indicate that in the original sample, the odds for mortality in six months for residents deemed unable to eat, were 2.56 to 1. After weighting the sample so that confounded effects of age, gender, and other disabilities were removed, the odds of mortality were reduced to 1.86 to 1.

The structured query language (SQL) code for completing the analysis of impact of unable to eat on survival using stratified Covariate Balancing is

/\*\*\*\*\*\* Analysis of Residents' Disabilities \*\*\*\*\*\*/

USE Assessment

DROP TABLE #Data, #Data2, #Data3, #Dead, #uEat, #MinAge, #Cases, #Controls, #Match

SELECT [Column 0] AS ID

,Cast([Column 1] AS Float) + CAST([Column 5] AS Float)/365. AS [Age]

,Iif ([Column 2]='M', 1, 0) AS [Male]

,[Column 3] AS [Number of assessments completed]

,[Column 4] AS [Days followed]

,CAST([Column 5] AS Float) AS [Days since first assessment]

,[Column 6] AS [days to last assessment]

,[Column 7] AS [Unable to eat]

,[Column 8] AS [unable to transfer]

,[Column 9] AS [Unable to groom]

,[Column 10] AS [Unable to toilet]

,[Column 11] AS [Unable to bathe]

,[Column 12] AS [Unable to walk]

,[Column 13] AS [Unable to dress]

,[Column 14] AS [Unable to bowel]

,[Column 15] AS [Unable to urine]

,[Column 16] AS [Dead]

,[Column 17] AS [Assessment number]

INTO #Data

FROM [Assessment].[dbo].[Assessments]

-- (1,306,456 row(s) affected)

-- Calculate Dead in 6 months

SELECT ID AS [ID Dead]

, [Days since first assessment] AS [Dead on day]

INTO #Dead FROM #Data WHERE Dead=1 -- (196,533 row(s) affected)

SELECT #Data.\*

, iif(#dead.[Dead on Day]-#Data.[Days since first assessment]

between 0 and 183, 1,0) AS [Dead6M]

INTO #Data2 FROM #Data left join #Dead ON [ID]=[ID Dead]

-- Drop Events after unable to eat

SELECT ID AS [ID unable to eat]

, Min([Days since first assessment]) AS [Unable to eat on day]

INTO #uEat FROM #Data

WHERE [Unable to eat]=1

GROUP BY [ID] -- Removes duplicates

-- (173067 row(s) affected)

Select MIN(Age) AS MinAge INTO #MinAge FROM #Data2 GROUP BY ID

Declare @AvgAge AS Float

SET @AvgAge=(SELECT Avg(MinAge) FROM #MinAge)

SELECT ID, [Assessment number], Dead6M, [Unable to Eat]

, Concat(IIF(Age<@AvgAge,'Y', 'O'), IIF(Male=1, 'M','F'),

IIF([unable to transfer]=1,'S',''), IIF([Unable to groom]=1, 'G',''),

IIF([Unable to toilet]=1, 'T',''), IIF([Unable to bathe]=1, 'B',''),

IIF([Unable to walk]=1, 'W', ''), IIF([Unable to dress]=1, 'D',''),

IIF([Unable to bowel]=1, 'L',''), IIF([Unable to urine]=1,'U','')) AS Strata

INTO #Data3 FROM #Data2 left join #uEat ON [ID]=[ID unable to eat]

WHERE #Data2.[Days since first assessment]<=#uEat.[Unable to eat on Day]

or #uEat.[Unable to eat on Day] is null

-- (1,087,271 row(s) affected)

-- Cases describe residents who are unable to eat

SELECT COUNT([ID]) AS nCases -- Number of assessment unable to eat

, Sum(IIF([Dead6M] = 1, 1., 0.)) AS DeadCase -- Number unable to eat and dead in 6 months

, SUM(IIF([Dead6M] = 0, 1., 0.)) AS AliveCase -- Number unable to eat and alive

, Strata

INTO #Cases

FROM #Data3

WHERE [Unable to Eat] = 1 -- Select only assessments unable to eat

GROUP BY Strata

--Controls are residents who are able to eat.

SELECT COUNT([ID]) AS nControls -- Number of assessment able to eat

, Sum(IIF([Dead6M] = 1, 1., 0.)) AS DeadControl -- Number able to eat and dead in 6 months

, SUM(IIF([Dead6M] = 0, 1., 0.)) AS AliveControl -- Number able to eat and alive

, Strata

INTO #Controls

FROM #Data3

WHERE [Unable to Eat] = 0 -- Select only assessments able to eat

GROUP BY Strata

-- Match cases with controls

SELECT nControls, DeadControl, AliveControl, #Cases.\*

INTO #Match

FROM #Cases inner join #Controls

ON #Cases.Strata =#Controls.Strata

-- Calculate common odds ratio

SELECT sum(DeadCase\*AliveControl/(DeadCase+AliveCase+DeadControl+AliveControl))/

sum(AliveCase\*DeadControl/(DeadCase+AliveCase+DeadControl+AliveControl)) As [Common Odds Ratio]

FROM #Match

-- Calculate overlap

Declare @TotalCases Float

SET @totalCases = (SELECT Sum(nCases) FROM #Cases)

SELECT ROUND(SUM(nCases)\*100/@TotalCases,2) as [Percent Overlap]

FROM #Match

## **[H1] Overlap Problem and Solutions to It**

The problem with stratification is that as the number of covariates increases, fewer and fewer cases and controls match. Along with this increase, the number of cases per stratum decreases, and combinations of covariates become quite rare. In these circumstances, it is possible that a large portion of the cases may not have matching controls and therefore are not used, reducing the generalizability of the findings. The findings will still be accurate for cases that were matched but perhaps not valid for other types of cases. The extent to which cases and controls are matched is referred to as *overlap* and is calculated as

In this equation, *c* is an index to cases, is 1 when the case is matched to a control, and 0 otherwise. The parameter indicates the percent of covariates in the case that were matched to the controls. If all covariates were matched, then Note that the percent of overlap does not depend on controls that were not matched to a case. The intent of the analysis is to examine the effect of treated patients—thus, what matters is matching to the cases. Unmatched controls do not change the treatment effect and therefore can be ignored. When the percent of overlap is low (e.g., lower than 80 percent), findings cannot be generalized, as many cases are not matched to controls.

At end of the covariate balancing of residents of nursing homes, 164,003 out of 164,017 cases who were unable to eat were matched on all covariates, to 865,849 out of 875,063 controls who were able to eat. The percent of overlap was calculated as

In this example, the data were collected on a large number of patients and the number of covariates was relatively small, so most of the cases were matched to the controls. This may not be the situation in other data sets. At least three strategies are available for increasing the percent of overlap

1. **[H2] Partial Matches Using Expected Values**

In this method, percent of overlap is increased by matching on a smaller set of variables among the unmatched cases. The unmatched case is altered to the largest portion of the case that matches to at least one control. The matched covariates are referred to as the shared component. The outcome for the altered cases is set to the expected outcome for all cases that share the common component. For example, suppose a male, 70-year-old patient with walking disability has no match among the controls. The closet match we can find is male patients with walking disabilities. As a result, the patient’s age is dropped from the analysis. The outcome for the new case is the average for all male disabled cases, which includes individuals in different age groups. This new case is matched to male patients unable to walk among the controls. The percent of overlap is improved in these cases by the portion of covariates matched, in this case by two out of three.

**[H2] Partial Matches Using Parents in Markov Blanket of Treatment**

A number of studies have shown that the Markov blanket can be used to reduce high-dimensional data to its essential variables see Tan and Liu 2013). The Markov blanket of treatment is a set of covariates that block the effect of other covariates on treatment (see chapter 20). Markov blankets include parents and co-parents (direct causes) and children (effects). Parents in the Markov Blanket are identified by focusing the analysis on independent variables that occur prior to treatment.

Algorithms for identifying the parents in a Markov blanket are discussed in chapter 20; here we focus on an algorithms that uses least absolute shrinkage and selection operator (LASSO) regression. LASSO regression is a type of regression that limits variables that have a statistically significant impact to those that have a large effect size. Prior to conducting the regression, we exclude covariates that occur after treatment. This step removes covariates in the causal path from treatment to outcome. For example, complications of treatment are excluded from the list of independent variables.

Next, the treatment variable is regressed on independent variables that occur before the treatment—for example, patient demographics, medical history, or comorbidities. Parents in the Markov blanket consist of covariates that (a) have a statistically significant impact on the outcome and (b) have an effect size greater than a preset cutoff value. A separate regression is done to verify that the covariates in the Markov blanket do not interact with any remaining covariates to have a statistically significant and large effect on treatment (more details of this procedure is presented in Chapter 20). Once the parents in the Markov blanket of treatment have been identified and verified, then remaining covariates are ignored. The procedure allows for matching to relevant and ignoring irrelevant covariates. The percent of overlap is now calculated among a smaller set of covariates and will be higher than when all covariates were balanced.

**[H2]** **New Synthetic Controls**

In this method, synthetic controls are created that would match to the features of currently unmatched cases. The method is influenced by the procedures used by artificial intelligence analysts to oversample underrepresented parts of the data (Chawla et al. 2002). First, the analysis is not done on all of the data. It focuses on control patients only, because the missing controls must reflect the pattern of outcomes among the controls. Second, working solely with control patients, the outcome is modelled as a function of the covariates (usually using regression or the two nearest cases), making sure that interactions among the covariates are included in the model. Last, the model is used to predict the outcome for the missing control cases by predicting the model using the parameters (covariates) of the unmatched cases.

For example, suppose that a male, 70-year-old resident who is unable to walk is an unmatched case. We need to look for a control that would match this case. Assume that for control patients, the outcome (survival rate) is predicted by the equation

Then, the missing control has the same feature as the unmatched case but is predicted to have the following outcome

Now that the predicted outcome is available, the predicted outcome can be added to missing controls and the analysis repeated with the synthetic controls matching to previously unmatched cases.

## **[H2] Example of Partial Stratification for Disabled Residents**

Exhibit 16.5 provides the estimated odds when one of the covariates is left unmatched. The first row shows the estimated unconfounded odds ratio using all covariates. The remaining rows remove one covariate at a time. For example, the exclusion of “unable to urinate” increases the number of cases matched to 100 percent. At the same time, the exclusion of “unable to urinate” does not change the estimated odds of mortality by much. After dropping any of the covariates, the odds of mortality for residents who are unable to eat ranged from 1.79 to 1.87, so the conclusion that residents who are unable to feed themselves are at increased risk of mortality does not change. This sensitivity analysis shows two things: (1) partial matches can increase the number of cases matched, and (2) in this case, going from exact to partial match had a negligible effect on the conclusions of the analysis.

**Exhibit 16.5** Sensitivity of Odds of Mortality for Residents Unable to Eat

|  |  |  |  |
| --- | --- | --- | --- |
| **Covariate Removed** | **Cases Unable to Eat**  ***n* (%)** | **Controls**  **Able to Eat**  ***n* (%)** | **Odds of Mortality** |
| None | 164,003 (99.9) | 865,849 (98.9) | 1.86 |
| Age | 164,009 (99.9) | 868,818 (99.2) | 1.87 |
| Gender | 164,016 (99.9) | 873,160 (99.7) | 1.85 |
| Unable to Bathe | 164,003 (99.9) | 865,849 (98.9) | 1.86 |
| Unable to Walk | 164,009 (99.9) | 868,818 (99.2) | 1.87 |
| Unable to Dress | 164,016 (99.9) | 873,160 (99.7) | 1.85 |
| Unable to Bowel | 164,017 (100) | 873,954 (99.8) | 1.79 |
| Unable to Urinate | 164,017 (100) | 874,624 (99.9) | 1.81 |
| Unable to Groom | 164,017 (100) | 874,624 (99.9) | 1.81 |
| Unable to Toilet | 164,017 (100) | 875,063 (100) | 1.83 |
| Unable to Sit | 164,017 (100) | 873,954 (99.8) | 1.79 |

## 

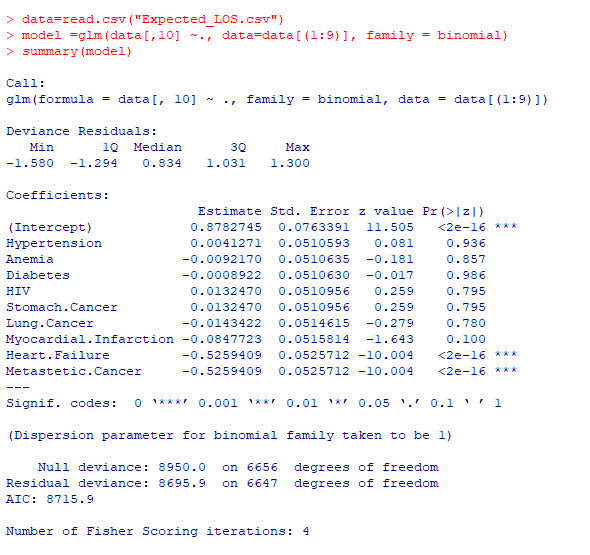
## **[H2] Example of Partial Stratification Using Parents in a Markov Blanket**

# In the data presented in question 1 of the problems in this chapter, we present length of stay of patients under the care of Dr. Smith and his peer group. These patients are described by ten comorbidities, making 210 = 1,024 distinct possible combinations. However, Dr. Smith and his group only care for 4,077 patients, raising the possibility that many cases may not have a matched control or, if matched, they would have few controls to estimate the effect of Dr. Smith on length of stay accurately. To remedy this problem with overlap, we can identify parents in the Markov blanket of Dr. Smith.

# We do so by first excluding all events that occur after treatment from the analysis. This step is relatively easy, as medical history is always collected prior to care delivered by Dr. Smith. The reported diagnoses are not complication of care but history that existed prior to care. Second, we regress being cared for by Dr. Smith on the ten diagnoses. The resulting regression is given in exhibit 16.6.

# Only heart failure and metastatic cancer have statistically significant relationship with care delivered by Dr. Smith. Furthermore, the effect of these two variables on the selection of Dr. Smith is large. The magnitude of the coefficient for both covariates is −0.52. In large data, I consider coefficients smaller than −.4 or larger than .4 to be large effects. Therefore, heart failure and metastatic cancer are parents in Dr. Smith’s Markov blanket. If we stratify these two covariates, all of the remaining eight diagnoses in the medical history are irrelevant. The analysis can be substantially simplified and percent of overlap increased by including only these two variables in the GROUP BY statement. Instead of examining the data in the combinations of ten covariates, we need only examine the combination of two, resulting in a higher number of patients falling into each stratum.

**Exhibit 16.6** Influence of Medical History on Selection of Dr. Smith



# [H1] Automated Removal of Confounding

Consider for a moment the difficulty of a computer arranging possible medication lists in EHRs to reflect the likely effectiveness of these medications. The computer, without the help of the analyst, has to assess the effectiveness of each medication. In removing the confounding, the computer must balance a large number of covariates, including demographics, medical history, mental health history, prior medication use, current diagnoses, allergic reactions, and, if available, genetic profile. This section provides SQL code for removing confounding via stratified covariate balancing.

In the initial step, all possible combinations of covariates are examined. In the computing literature, this is referred to as an *item set*. In addition, for each item set, the count of occurrence of the combination is reported. The following SQL code can be used to identify combinations of a small number of covariates, and faster algorithms are available for identifying larger set of covariates.

SELECT [Antidepressant]

, [Medical History]

, [Mental Health History]

, [Medication History]

, [Current Allergies]

, [Current Diagnoses]

, COUNT(Id) AS [Number of Patients]

INTO #Combinations  
FROM dbo.Data  
GROUP BY [Antidepressant], [Medical History], [Mental Health History] , [Medication History, [Current Allergies], [Current Diagnoses]

In the next step, the combinations of the covariates are used to define the stratum, and within each stratum, the code divides patients into those who received the treatment (referred to as *cases*) and those who did not (referred to as *controls*).

SELECT [Number of Patients] AS Cases, #Combinations.\*

INTO #Cases

FROM #Combinations

WHERE [Antidepressant] =1;

Controls are patients without the treatment (those who did not receive the antidepressant in the previous code). The following SQL code was used to identify the controls among patients who did not receive the antidepressant.

SELECT [Number of Patients] AS Controls, #Combinations.\*

INTO #Controls

FROM #Combinations

WHERE [Antidepressant]=0;

Next,cases and controls are matched together using the INNER JOIN statement for connecting these two tables. In addition, the weight for controls is calculated as the ratio of cases to controls. The following is the SQL code used to match the two tables.

SELECT

[Cases]/[Controls] AS Weight

, #Controls.\*

, #Cases.\*

INTO #Match

FROM #Cases INNER JOIN #Controls

ON

#Cases.[Medical History] = #Controls.[Medical History]

, #Cases.[Mental Health History] = #Controls.[Mental Health History]

, #Cases.[Medication History] = #Controls.[Medication History]

, #Cases.[Current Allergies] = #Controls.[Current Allergies]

, #Cases.[Current Diagnoses] = #Controls.[Current Allergies]

The set of weights estimated in the previous step guarantees that covariates are balanced. Some clinicians would like to see that this is true. The computer can reassure them by providing rates of covariates among treated (cases) and untreated groups (controls). The following SQL shows the rate of medical history covariates after the use of weights.

SELECT

Sum(IIF(Antidepressant =1,1,Weight)\*[Medical History]) as [Weighted Number with Medical History]

, [Antidepressant]

INTO #WeightedMedicalHistory

FROM #Match INNER JOIN Data

ON

[Cases]![Medical History] = Controls![Medical History]

, [Cases]![Mental Health History] = [Controls]![Mental Health History]

, [Cases]![Medication History] = [Controls]![Medication History]

, [Cases]![Current Allergies] = [Controls]![Current Allergies]

, [Cases]![Current Diagnoses] = [Controls]![Current Allergies]

GROUP BY [Antidepressant]

SELECT

Sum([Medical History]) as [Number with Medical History]

, [Antidepressant]

INTO #MedicalHistory

FROM #Match INNER JOIN dbo.Data

ON

[Cases]![Medical History] = Controls![Medical History]

, [Cases]![Mental Health History] = [Controls]![Mental Health History]

, [Cases]![Medication History] = [Controls]![Medication History]

, [Cases]![Current Allergies] = [Controls]![Current Allergies]

, [Cases]![Current Diagnoses] = [Controls]![Current Allergies]

GROUP BY [Antidepressant]

Similar analysis can be done for all other covariates. When a computer reminds clinicians about the comparative effectiveness of a medication, it is important to reassure them that all covariates are balanced. One way is to show that the number of patients with a particular medical history was different before balancing the data but the same after.The difference between treated and untreated groups is calculated and displayed. If the outcome variable is binary, this difference is calculated as an odds ratio. If the outcome variable is continuous, this difference is calculated as a weighted average. The following SQL code shows the calculation for a continuous outcome.

SELECT

AVG(IIF(Antidepressant =1,1,Weight)\*Outcome) as [Average Outcome]

, [Antidepressant]

INTO #Impact

FROM #Match INNER JOIN dbo.Data

ON

[Cases]![Medical History] = Controls![Medical History]

, [Cases]![Mental Health History] = [Controls]![Mental Health History]

, [Cases]![Medication History] = [Controls]![Medication History]

, [Cases]![Current Allergies] = [Controls]![Current Allergies]

, [Cases]![Current Diagnoses] = [Controls]![Current Allergies]

GROUP BY [Antidepressant]

Sometimes too many covariates are available and too few cases fall in the various strata. To increase the number of controls that match the cases, the computer can progressively drop one covariate and redo the analysis. By dropping a variable, the computer is replacing the exact match to all of the patient’s characteristics with partial matching. If the dropped variable does not change the order of comparative effectiveness of the medications, then the procedure is not sensitive to the variable, and the partial matching makes sense.

# [H2] R Package

To assist your use of stratified covariate balancing, a free R package is available online (search for “StratifiedBalancing” with “R Package” on the web).

# [H1] Summary

The key message of this chapter is that confounding can be removed and covariates balanced through stratification. The process provides weights for balancing continuous outcomes. These weights replace propensity scoring and are relatively easier to obtain, especially in EHRs. The procedures described in this chapter allow the analyst to balance data in EHRs through SQL, without reliance on statistical packages.

# [H1] Supplemental Resources

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

**[H1] References**

Alemi, F., A. ElRafey, and I. Avramovic. 2018. “Covariate Balancing Through Naturally Occurring Strata.” *Health Services Research* 53 (1): 273–92.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeye. 2002. “SMOTE: Synthetic Minority Over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16 (1): 321–57.

Levy, C. R., M. Zargoush, A. E. Williams AE, A. R.Williams, P. Giang, J. Wojtusiak, R. E. Kheirbek, and F. Alemi. 2016. “Sequence of Functional Loss and Recovery in Nursing Homes.” *Gerontologist* 56 (1): 52–61.

Rosenbaum, P. R., and D. B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70: 41–55.

Tan, Y., and Z. Liu. 2013. “Feature Selection and Prediction with a Markov Blanket Structure Learning Algorithm.” *BMC Bioinformatics* 14 (Suppl 17): A3.