Confounding is one of the major concerns regarding the validity of the results from observational studies. Epidemiological methods to adjust for confounding, such as multivariable adjustment using logistic regression and propensity scores based methods, may have problems trying to balance exposure groups with respect to the potential confounders. A new reweighting method called entropy balancing can achieve perfect balance of the exposure groups on the covariates with respect to their means (for continuous variables) and proportions (for categorical variables).

**Logistic Regression Model:**

The logistic regression model is a statistical model used to predict a dichotomous response using predictors. In our analysis, we will use logistic regression to assess the relationship using predictors. In our analysis, we will use logistic regression to assess the relationship using predictors. In our analysis, we will use logistic regression to assess the relationship using predictors.

The logistic regression model is a statistical model used to predict a dichotomous response using predictors. In our analysis, we will use logistic regression to assess the relationship using predictors.

**Methods**

**Two-level Treatment:**

We balance the two groups by reweighting each group separately in such a way as to have the same mean as the whole sample. This approach allows us to be able to estimate the population average treatment effect (PATE).

The constraints can be described mathematically as follows:

\[ \sum_{i \in I_D} w_i Y_i = \sum_{i \in I_T} w_i Y_i = n_T \]

\[ \sum_{i \in I_D} w_i = 1, \quad j \in D, \quad w_i \geq 0 \text{ for all } i \]

The first formula indicates that after applying the weights to each unit, both the control group (D=0) and the treatment group (D=1) have the first moment function (mean) of each covariate equal to the first moment function (mean) for that covariate for the whole sample. The second formula indicates that the weights of each group (treatment and control) should sum up to 1. The third formula indicates that all the weights should be non-negative.

To preserve the original sample sizes for the treatment and control groups, we further rescale the weights previously obtained:

\[ \sum_{i \in I_D} w_i = n_D, \quad j \in D, \quad w_i \geq 0 \text{ for all } i \]

**Three-level Treatment:**

We extended the scope of the entropy balancing method to the situation of a three-level exposure. Each one of the three groups of subjects are balanced by constraining them such that each of the groups’ covariate means match the covariate means of the whole dataset:

\[ \sum_{i \in I_D} w_i Y_i = \sum_{i \in I_T} w_i Y_i = \sum_{i \in I_M} w_i Y_i = n_T \]

Again, the weights are rescaled so that they will add up to each group’s original sample size:

\[ \sum_{i \in I_D} w_i = n_D, \quad \sum_{i \in I_T} w_i = n_T, \quad \sum_{i \in I_M} w_i = n_M \]

The weights are obtained by using a modified version of the ebal package from R.

**Weighted Logistic Regression Analyses:**

Odds ratio and 95% CI are calculated using logistic regression and survey logistic regression in R and in SAS. For the survey logistic regression models, where the entropy balancing weights were used as sampling weights, we use PROC SURVEYLOGISTIC from SAS and the survey package from R.

**Results**

The data to analyze is from the Baltimore-Washington Infant Study (BWIS). This specific data has been previously analyzed for a practicum project using a logistic regression model by William R. Wisniewski.

The cases (n=845) were defined as all live born infants with a Ventricular Septal Defect (VSD), while the control infants (n=3,591) were defined as any infant without any CVM. A logistic regression model was built using VSD as the binary outcome variable and BMI group as the main exposure variable. Confounders included in the model were maternal age at conception, infant race, maternal education, maternal smoking (cigs/day), paternal smoking (cigs/day), maternal frequency alcohol use, family CVM history, previous premature births, maternal diabetes, and previous miscarriages.

**Logistic Regression and Entropy Balancing for a Three-Level Main Exposure Variable**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Logistic Regression</th>
<th>Entropy Balance &amp; Logistic Regression</th>
<th>Entropy Balance &amp; Survey Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI (kg/m²)</td>
<td>Underweight (&lt; 18.5)</td>
<td>0.71 (0.53-0.93)</td>
<td>0.70 (0.52-0.94)</td>
</tr>
<tr>
<td></td>
<td>Normal Weight (18.5-24.9)</td>
<td>reference</td>
<td>reference</td>
</tr>
<tr>
<td></td>
<td>Overweight (≥ 25.0)</td>
<td>0.82 (0.68-0.98)</td>
<td>0.84 (0.70-1.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.84 (0.70-1.02)</td>
</tr>
</tbody>
</table>

The results (Odds Ratio and 95% CI) obtained from the reweighted data are very close to the results from the usual logistic regression model adjusting for confounders. For this specific dataset the entropy balancing method worked as well as the logistic regression model.

**Conclusion**

In this practicum project we extended the scope of the entropy balancing method from the case involving two-level exposure variables to that of three-level exposure variables. Using the entropy balancing method we are able to achieve perfect balance of the potential confounders. Thus, we recommend that the entropy balancing method should be used as a sensitivity analysis method.

A comparison of propensity score based methods and the entropy balancing method on additional three-level exposure data and on simulated data will be considered as future work to assess the performance of the entropy balancing method.