Estimating the Causal Effect of Low Birth Weight on Infant Mortality

Katherine Grace Wasmer **Vaibhava Lakshmi Ravidashek**

Abstract

This study estimates the causal effect of low birth weight (LBW) on infant mortality using R-Learner and Double Machine Learning (DML) frameworks, analyzing same-sex twin data from the 2023 National Center for Health Statistics. The R-Learner model yielded a Conditional Average Treatment Effect (CATE) of 0.0098, while the DML model produced a slightly higher CATE of 0.013. The R-Learner's selective pre-processing enhanced interpretability and computational efficiency, whereas the DML's integration of SHapley Additive exPlanations (SHAP) provided deeper insights into feature impacts. Both models highlight LBW's significant risk to infant mortality and demonstrate the robustness of advanced machine learning techniques in causal inference, with DML offering slightly more comprehensive results. This research underscores the importance of addressing LBW in public health policies.

1 Introduction

Infant mortality is a prominent issue in the public health domain that lends itself to numerous directions of research on the topic. In the United States, the Center for Disease Control and Prevention (CDC) estimates that roughly 0.56 percent of babies do not survive past the neonatal stage (Driscol & Ely, 2024). The CDC's National Center for Health Statistics (NCHS) reports these rates on a quarterly basis and produces final reports at the end of each year, which allows us to identify trends over time. To understand the significance of the current statistics, it's important to compare them to historical rates.

As a point of reference, roughly 3 percent of infants did not survive past the first year in 1950. By the turn of the century, however, this number was nearly quartered at 0.71 percent (Field et al., Chapter 2). This means that over the past two decades, the infant mortality rate in the United States has decreased by twenty-five percent. What are the reasons for this improvement, and why has the curve flattened in recent years?

Revolutionary medical technology and increased awareness of disease prevention are two major determinants in this shift. Most noteworthy of these developments was the polio vaccine; in 1955, polio was feared almost as much as the atomic bomb (Fitzpatrick, 2006), but it is now standard protocol to vaccinate infants against this disease, which has become nearly nonexistent in the United States as a result. Moreover, the emergence of environmental advocacy in the 1970s led to a rise in anti-tobacco campaigns. People became more aware of the dangers of secondhand smoke on an unborn baby, and mothers were discouraged from smoking while pregnant. By the 1990s, it was common knowledge that tobacco had a negative impact on developing babies, and smoking was not as normalized in the media as it had been fifty years prior (*The Rise of Anti-Smoking Movements · Yale University Library Online Exhibitions*, n.d.).

These two examples accentuate the decrease of infant deaths caused by infectious disease. Consequently, the leading causes of contemporary infant mortality are often based in genetic or congenital

Preprint. Under review.

factors, which are far more difficult to treat. Since 1999, low birth weight (**LBW**) is the one of the most common causes of infant deaths in the U.S., second only to birth defects (*Infant Deaths*, 2024).

Our research question explores the impact of LBW on survival outcomes of infants. The choice of LBW as the **intervention** (also known as the **treatment effect**) in our experiment aligns with previous literature on the topic. Almond et al. emphasizes the ramifications of LBW on overall quality of life and well-being. Infants with a lower birth weight have a higher likelihood of dying within the first 12 months of life. Even if they do survive, they are more susceptible to increased hospital visits and costs, developmental delays, and health problems. [1]

Furthermore, the LBW does not have a single definitive cause, which makes it a useful treatment effect. Our machine learning pipeline conditions on multiple features that enable us to infer how infant mortality effects different demographics. The most recent CDC statistics on the topic indicate racial disparities, particularly within Black and indigenous communities. By implementing orthogonal machine learning, we analyze numerous determinants (including race, the mother's age, her education status, and social class) and their relationship with the treatment effect and mortality outcome.

2 Methodology & Contributions

2.1 Identifying the Estimand

Recent advancements in orthogonal machine learning methodologies provide powerful tools for estimating the conditional average treatment effect (CATE), which is the target estimand in our research. The core advantage of the CATE lies in its ability to integrate insights from both experimental and observational data. The combination of these two types of data are crucial for developing robust models, especially in the healthcare setting where diverse data sources are common.

In rudimentary terms, the CATE quantifies the expected difference in outcomes between treated and untreated groups, conditional on a set of covariates. It serves as a cornerstone for understanding how treatment effects vary across different subgroups within a population.

Formally, the CATE is defined as

$$\tau^*(X) = E[Y(T=1) \mid X] - E[Y(T=0) \mid X] \tag{1}$$

where:

- T: A binary variable that determines whether or not the individual received the treatment.
- $Y(T=1)$: The potential outcome if the individual receives the treatment.
- $Y(T=0)$: The potential outcome if the individual does not receive the treatment.
- X: A vector of covariates or features that characterize the individual.

In our applied research, $T = 0$ represents the subset of infants with a birth weight below a fixed threshold, and $T = 1$ represents the infants with a birth weight above this specified threshold. Therefore, $Y(0)$ and $Y(1)$ represent the aggregate of potential survival outcomes for infants in each treatment group. It is crucial to highlight that $Y(0)$ and $Y(1)$ (and $T = 0$ and $T = 1$, by extension) are mutually exclusive categories. In other words, a baby cannot be both below and above the weight cutoff.

This also means that it is impossible to directly compute $\tau^*(X)$, so the estimator $\hat{\tau}(X)$ is calculated instead. In the scope of epidemiology and medical research, twin studies are particularly useful for minimizing $\tau^*(X) - \hat{\tau}(X)$. By narrowing the sample size down to same-sex twins who had all covariates in common, but opposite treatment effects, we treated each twin pair as the "same person with different outcomes". Our study thus operates under the assumption that $\hat{\tau}(X) \approx \tau^*(X)$.

The procedure below describes in detail how we extracted the data to simulate a case where both potential outcomes are visible.

1. We utilized the 2023 Live Birth dataset from the NCHS, which is the most recent complete record of births in the United States. The pandas library in Python was used to filter the data to only a plurality status of "twin". After verifying that the rows of the data set were listed in chronological order by date of birth, we assigned a pair ID to each set of twins. The code

for creating these IDs was borrowed and modified from the MIT Health Lab, where similar research on live birth data has been conducted.

2. After deriving these pair IDs, we calculated the descriptive statistics for the weight variable on the twin data. To eliminate any potential bias, we randomly sampled a subset of n = 1,210 twins on which to perform these statistics. For this subset, we calculated a median weight of 2,400 ounces, and a mean of 2,341 ounces. Due to the skewness of the data, we established 2,400 as the cutoff for which treatment group the infants would be placed in. This sample median is close enough to the median of the entire set of same-sex twins. Our rationale for computing the descriptive statistics on only the twins (and not the entire 2023 Live Births dataset) is that twins have a higher propensity for a LBW than singleton babies do. We wanted to create a cutoff that accurately reflected this discrepancy.

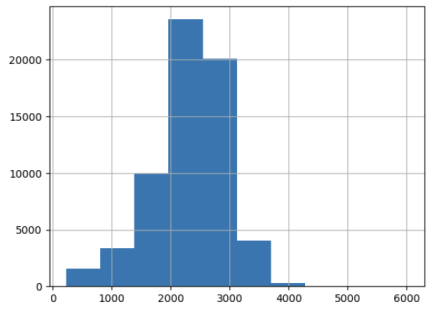

Figure 1: A histogram of 605 randomly selected twin pairs (n = 1,210), indicating a minor right skew in the weight distribution.

3. After establishing $T = 0$ for any infant below 2,400 ounces and $T = 1$ for any infant above that weight, we created a new column called "treatment_effect" and assigned individuals to each group accordingly. We grouped the data frame by the "pair_ID" variable and subsetted pairs that were both male or both female in order to keep the X covariates identical for each set of twins. Finally, we reduced the data only to twins who had different treatment effects, resulting in a final sample size of n = 21,366.

The procedure for data cleaning and handling certifies that the key assumptions of the CATE are satisfied. Although the CATE can be estimated with machine learning algorithms that capture complex relationships between covariates and outcomes, one should be familiar with the mathematical assumptions that define this particular estimand and how they apply to our study.

1. **SUTVA:**
   (i) The consistency component of Stable Unit-Treatment Value Assumption (SUTVA) states that $Y(a) = Y$ if $A = a$. In our project, Y(0) will always reflect the outcome of infants in group $T = 0$, and Y(1) will always reflect the outcome of infants in group $T = 1$. This ensures a clear separation between the two treatment groups and their potential outcomes. (Chernozhukov et. al, Ch. 2, 2024, pg. 44).
   (ii) The no interference component of the SUTVA indicates that $Y_i(a_1, ..., a_i, ...a_n) = Y_i(a_i)$ (Makar, 2024). For any individual in the data set, their treatment status does not depend on any other infant's $T = t \in (0, 1)$. Assigning T based on the fixed threshold of 2,400 ounces allows for this no interference property.
2. **Ignorability**: The treatment assignment is independent of the potential outcomes, given the covariates X. Mathematically, the ignorability property is written as $(Y(1), Y(0) \perp T \mid X)$. Accordingly, the distributions of $Y(0), Y(1)$ depend on the predictors X, instead of T. (Chernozhukov et. al, Ch. 5, 2024, pg. 132).
3. **Overlap**: Each individual has a positive probability of receiving the treatment given the covariates $(0 < P(T = 1 \mid X) < 1)$. In other words, the distribution of each covariate should be roughly equal in both treatment groups. One example in the full Live Births data set is the time of prenatal care (denoted as "pcare5"); the mean for $T = 0$ is 1.34 trimesters,

and the mean for $T = 1$ is 1.42 trimesters, so these values are roughly equal and likely follow similar distributions.

2.2 Definition of R-Learner

The R-Learner framework, introduced by *Nie and Wager (2020)*, provides a flexible and robust approach to estimating heterogeneous treatment effects. Unlike traditional methods, the R-Learner separates the estimation of baseline outcomes from the estimation of treatment effects, which enhances its adaptability to complex, high-dimensional data.

The R-Learner proceeds in the following steps:

1. **Residualize the Outcome and Treatment**: $\tilde{Y} = Y - g(X)$,
   $\tilde{T} = T - e(X)$, where $g(X) = \mathbb{E}[Y \mid X]$ is the baseline outcome model, and $e(X) = P(T = 1 \mid X)$ is the propensity score model. These residuals remove the influence of covariates on both the outcome and treatment.
2. **Solve the Residualized Regression Problem**: The treatment effect $\tau(X)$ is then estimated by solving: $\hat{\tau}(X) = \arg\min_\tau \mathbb{E}\left[\left(\tilde{Y} - \tilde{T} \cdot \tau(X)\right)^2\right]$, which focuses purely on the relationship between the residualized treatment and outcome.

The key advantage of the R-Learner is its modular structure, allowing for the use of any machine learning models to estimate $g(X)$, $e(X)$, and $\tau(X)$. This flexibility makes it particularly suitable for high-dimensional and complex data.

2.3 Definition of DML

Double Machine Learning (DML), developed by *Chernozhukov et al. (2018)*, is a powerful method for estimating treatment effects while accounting for high-dimensional confounders. DML employs a two-stage process that leverages cross-fitting to reduce overfitting and enhance robustness.

The DML approach involves the following steps:

1. **Model Outcome and Treatment with Cross-Fitting**: Use machine learning models to estimate the conditional mean of the outcome $g(X) = \mathbb{E}[Y \mid X]$ and the propensity score $e(X) = P(T = 1 \mid X)$. Cross-fitting is employed, where data is split into folds, and models trained on one fold are used to make predictions on another. This prevents overfitting.
2. **Compute Residuals**: Calculate the residuals for the outcome and treatment: $\tilde{Y} = Y - \hat{g}(X)$,
   $\tilde{T} = T - \hat{e}(X)$. These residuals remove the effects of covariates, isolating the relationship between treatment and outcome.
3. **Estimate the Treatment Effect**: Solve a partially linear regression model using the residuals: $\tilde{Y} = \tau \cdot \tilde{T} + \epsilon$, where τ is the treatment effect of interest, and ϵ is the error term.

DML provides asymptotically unbiased estimates of treatment effects and handles high-dimensional data effectively. By combining modern machine learning techniques with classical econometric principles, it offers a rigorous framework for causal inference in complex settings.

3 Findings & Discussion

We first estimated the CATE with the R-learner approach due to its straightforward implementation. As mentioned in the Methodology section, it is quite simple to estimate the CATE with machine learning algorithms; we chose to integrate EconML with Scikit-learn for our computation. Our rationale for these Python libraries was to easily perform orthogonal and double machine learning while using regressors that had native support of null values in X. For this part of the study, we used Scikit-learn's HistGradientBoostingRegressor within the R-learner. In **gradient boosting**, a loss function is optimized on an ensemble of weak learners (*Gradient Boosting*, 2019), thus generating a strong learner derived from an additive model of weak learners. The HistGradientBoostingRegressor

is advantageous for data sets with more than 10,000 rows, and there are a wide variety of loss functions that can be used.

The original CATE with this regressor and framework comprised 24 covariates surrounding the infancy; these covariates are listed as follows: sex, mother's age (divided into groups of 5 years, denoted as "mager9" in the data set), birth weight, the number of siblings who had died during infancy (denoted as "priordead", the number of abortions the mother had during previous pregnancies (denoted as "priorterm"), the facility of the birth, the trimester in which the mother started receiving prenatal care ("precare5"), the mother's nativity status to the U.S., the mother's residency status, the mother's race, whether or not paternity was acknowledged, the marital status of the biological parents, the mother's education level, whether or not the mother died during childbirth, smoking during the first, second, and third trimesters (all separate variables), pre-pregnancy diabetes, gestational diabetes, hypertension eclampsia, the STI status during pregnancy, whether or not infertility treatment was used, whether or not the mother used fertility-enhancing drugs, and the pair ID of the twins.

We fit the R-learner model with this data (including all null values, due to the native acceptance of missing data) and estimated the effect using the subset of data that did not contain any null values. The categorical columns were transformed with LabelEncoder. Initially, using non-imputed data and no additional encoding resulted in an estimated CATE of 0.0003, suggesting a small difference in survival outcome between the treatment groups. However, this statistic cannot be deemed reliable due to the limited sample size (n = 122); these were the only rows out of the entire data set didn't contain any null values.

To remedy the issue of the poor test size, we attempted to impute the entire data set, using the median values of each numerical column as the imputed value. We fed this imputed data set to the model and estimated its effect using this imputed data to obtain a CATE of 0.104. Although this value was a significant improvement over the previous one, it was biased for a variety of reasons. We realized that the LabelEncoder was problematic for categorical data with null values, because assigning them as 0 could lead to issues with the numerical columns and return an inaccurate result. Moreover, we dropped the "dbwt" column, because the infant's birth weight was already accounted for with the treatment effect. This variable also violated the **overlap assumption**, since the treatment effect depended on the birth weight. Finally, by imputing various percentages of the missing data,fitting them to the model, and estimating the effect with the set of 122 non-missing rows, the CATE significantly fluctuated, as depicted in Figure 2.

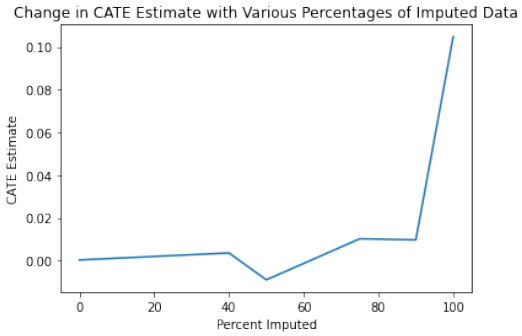

Figure 2: The CATE estimate based on the percentage of imputed null values in the data set.

In lieu of LabelEncoder, we changed our approach to handling categorical data (and missing data by extension) by utilizing OneHotEncoder. This produced a data set that contained no null values, but was not imputed. Instead, the null values were treated as features. For example, the variable "dmar" (the martial status of the parents) had responses "Yes", "No", and "NaN". With OneHotEncoder, this column transformed into 3 boolean columns that we respectively renamed "married_parents", "unmarried_parents", and "marriage_NaN". By fitting and estimating this new fully-encoded data frame, we obtained a CATE of 0.01. The remaining issue, however, was that the encoding resulted in thousands of columns due to the variety of numerical values. While the CATE estimate may be more accurate in this case, it was practically impossible to interpret the findings and determine which encoded column mapped to which original column.

To resolve these challenges, a refined pre-processing strategy was implemented; categorical columns were selectively one-hot encoded, numerical columns retained their null values, and the testing size expanded to 18,612 non-null rows. This approach yielded an estimated CATE of 0.0098, which can be rounded to 0.01, thus obtaining the same value as we did with the fully encoded data. However, this partially encoded approach significantly enhanced interpretability and reduced computational complexity. (Unless specified, please assume that any findings with the CATE were derived with this pre-processing technique.) Furthermore, the use of HistGradientBoostingRegressor across the model's stages ensured robust handling of missing values and effective modeling of complex feature interactions, solidifying the model's reliability. This idea was tested by sampling various subsets of the data to feed to the model and verifying that the CATE was roughly equal to 0.0098.

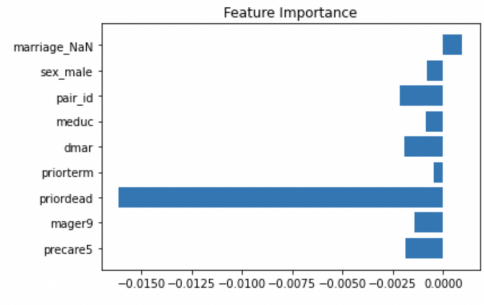

Figure 3: A bar graph showing the non-zero feature importance values for the corresponding predictors in X.

These findings emphasize the critical role of pre-processing in balancing bias, sample size, and interpretability in Causal Inference tasks. We explored the driving factors in outcome predictions with the R-learner model by developing a feature important plot (Fig. 3). At first glance, the "priordead" variable appears to exhibit a significant relationship between prior child deaths and whether or not the child survived infancy. However, the feature importance is only -0.015, which is a very weak correlation and cannot be used to make strong predictions. Although none of the features appear to heavily contribute to the outcome, it also illustrates the importance of having multiple covariates. On their own, these variables have minimal influence on the treatment effect, which does align with the science. LBW, like many medical conditions, is affected by genetic, environmental, and epigenetic factors. There is not one single cause, so we should not expect a single variable to significantly influence the outcome or the treatment effect.

A more practical alternative to studying each feature importance is to stratify different columns and determine if the estimated CATE varies across groups. This approach ties back into the importance of accounting for heterogeneity in medical data. We found that by studying different groups, some had little to no difference on the causal effect, while others had a larger impact. For this context, we define "different" as being 10 percent greater or 10 percent smaller than the crude CATE estimate of 0.0098. Any stratified value outside of $[0.0088, 0.0108]$ will be considered significant, and the feature will be accordingly labeled a confounder or effect modifier as appropriate.

male	0.0098
female	0.0098

< 15	0.0109
15-19	0.0105
20-24	0.0101
25-29	0.0097
30-34	0.0097
35-39	0.0098
40-44	0.0097
45-49	0.011

Table 1: The CATE stratified by sex (left), and the mother's age (right). The data shows that if the mother is younger than 15 years of age, or 45 or older, it modifies the causal effect of LBW on infant mortality, thus making it an **effect modifier**.

3.1 DML Results

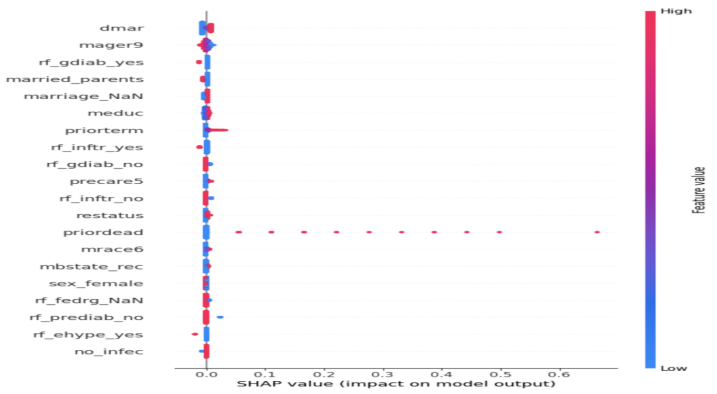

Figure 4: The SHAP summary plot for Double Machine Learning model outcomes.

The findings from the Double Machine Learning (DML) model demonstrate its effectiveness in estimating treatment effects by combining advanced pre-processing techniques with robust machine learning models. The DML framework utilized RandomForestRegressor as the base model for both the treatment (T) and outcome (Y) stages, leveraging its strength in capturing complex, non-linear relationships between features while handling datasets with mixed variable types. Numerical columns were preserved with missing values, while categorical variables were selectively one-hot encoded to maintain interpretability without unnecessarily inflating the feature space. Expanding the dataset to approximately 18,000 samples for testing significantly improved the reliability of the causal estimates, yielding an estimated Conditional Average Treatment Effect (CATE) of 0.013, consistent across various sub-samples.

male	0.013
female	0.012

Table 2: The CATE stratified by sex (DML model)

A notable addition to the interpretability of the model was the use of SHAP (SHapley Additive exPlanations) values, which provided insights into the contribution of individual features to the treatment effect.
The Shapley value for feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right]$$

where:

- ϕ_i: Shapley value for feature i.
- N: The set of all features.
- S: A subset of features not including i ($S \subseteq N \setminus \{i\}$).
- $|S|$: The size (number of features) of subset S.
- $|N|$: The total number of features.
- $v(S)$: The value function for subset S, representing the prediction or contribution of subset S.

By decomposing the predictions, SHAP values allowed for a transparent understanding of how specific variables influenced the model's outcomes, enhancing confidence in the results. For instance, from the Figure. 4, it is evident that features like "dmar" and "mager9" with higher SHAP values had a more pronounced impact on treatment estimates, guiding better interpretability of causal relationships. Whereas, features like "no_infec" and "rf_ehype_yes" have the least impact on treatment estimates.

4 Conclusion

In this study, we sought to estimate the causal effect of low birth weight (LBW) on infant mortality using advanced methodologies in causal inference, specifically the R-Learner and Double Machine Learning (DML) frameworks. The amalgamation of these two frameworks allowed for easier interpretability and provided pathways to determining effect modifiers and important features. Our comprehensive data processing and robust modeling efforts support the role of birth weight in infant survival outcomes. Our findings suggest that infants with low birth weights have a slightly heightened risk of mortality compared to their higher birth weight counterparts, reinforcing LBW as a crucial factor in public health strategies aimed at reducing infant mortality rates. The use of same-sex twin pairs as a natural experiment effectively mitigated potential confounding variables, allowing for more reliable estimations of the treatment effect.

The nuanced pre-processing strategies, including selective one-hot encoding and handling missing values, proved essential in obtaining interpretable and robust causal estimates. By experimenting with different imputation and encoding methods, we estimated a relatively unbiased CATE. The R-Learner estimated a Conditional Average Treatment Effect (CATE) of approximately 0.0098, while the DML approach yielded a slightly higher yet consistent CATE of 0.013. These results underline the importance of rigorous data handling and method selection in causal inference studies.

Additionally, the application of SHapley Additive exPlanations (SHAP) values in the DML framework provided further interpretability, highlighting the significant impact of specific co-variates such as prior child deaths and prenatal care on the estimated treatment effects. This level of interpretability is crucial for translating research findings into actionable public health policies. Our study highlights the value of integrating machine learning techniques with traditional causal inference, providing a robust framework for analyzing complex, high-dimensional healthcare data. Future research could expand on this foundation by exploring additional factors influencing LBW and employing more granular data to refine these estimations further.

References

[1] Almond, D., Chay, K. Y., & Lee, D. S. (2004). The Costs of Low Birth Weight. *National Bureau of Economic Research.* https://www.nber.org/papers/w10552

[2] Chernozhukov, Victor, et al. *Applied Causal Inference Powered by ML and AI*. Vol. 0.1.1, Online, 28 July 2024, causalml-book.org/.

[3] Fitzpatrick M. (2006). The Cutter Incident: How America's First Polio Vaccine Led to a Growing Vaccine Crisis. *Journal of the Royal Society of Medicine*, 99(3), 156.

[4] *Gradient Boosting.* (2019, May 17). DeepAI. https://deepai.org/machine-learning-glossary-and-terms/gradient-boosting

[5] *Infant Mortality*. (2024, September 16). U.S. Centers for Disease Control and Prevention. CDC - Infant Maternal Health. https://www.cdc.gov/maternal-infant-health/infant-mortality/index.html

[6] *The Rise of Anti-smoking Movements · Yale University Library Online Exhibitions.* (n.d.). Online-exhibits.library.yale.edu. https://onlineexhibits.library.yale.edu/s/sellingsmoke/page/antismoking

[7] Makar, M. (2024). *6 - Estimation under ignorability (intro)* [Powerpoint slides]. University of Michigan Computer Science and Engineering Canvas.

[8] Michael C Knaus, Double machine learning-based programme evaluation under unconfoundedness, The Econometrics Journal, Volume 25, Issue 3, September 2022, Pages 602–627, https://doi.org/10.1093/ectj/utac015

[9] Osterman, Michelle J.K. *User Guide to the 2023 Natality Public Use File* . Center for Disease Control and Prevention.

[10] Wu, L. amp; Yang, S.. (2022). Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. Proceedings of the First Conference on Causal Learning and Reasoning, in Proceedings of Machine Learning Research 177:904-926 Available from https://proceedings.mlr.press/v177/wu22a.html.

5 Task Overview

5.1 Contributions by Kate Wasmer

- Performed data cleaning and preprocessing for the Live Births 2023 dataset.
- Conducted random sampling of subsets to impute missing data and evaluate CATE (Conditional Average Treatment Effect) estimates.
- Applied One-Hot Encoding as detailed in the Results Findings section.
- Optimized the CATE calculation through iterative trial-and-error methods.

5.2 Contributions by Vaibhava Lakshmi Ravideshik

- Developed and implemented the code for R-Learner and Double Machine Learning (DML) frameworks.
- Conducted SHAP (SHapley Additive exPlanations) analysis to explain the predictions of the DML model.
- Performed an extensive literature review to support the methodologies used.
- Analyzed and highlighted feature importance values within the R-Learner framework.