Causal Inference (097400) - Project

Edan Kinderman, 316384445
Itamar Harel, 212362024

August 2024

1 Introduction

There are many factors influencing fertility rates. In this project, we investigate the effect of some subjective characteristics in youth on their fertility rate later in life. For the rest of this section, we expand on the causal question and choice of dataset. In Section 2 we describe the feature selection and definition process, and discuss shortcomings of the dataset as well as relevant data manipulations used to deal with them. In Section 3 we present some preliminary statistical properties of the data, followed by the causal analysis in Sections 4 and 5. We conclude with a brief discussion of our results in Section 6. [1]

1.1 Objective

In this project, we explore how a person's expectations and opinions about the number of children they will have in the future influence the actual number of children they have later in life. We address two specific causal questions, each with a different treatment but with the same outcome:

1. What is the causal effect of a person's **expected number of children**, as reported in their youth, on the actual number of children they eventually have?
2. What is the causal effect of a person's opinion on the **ideal number of children**, as expressed in their youth, on the actual number of children they eventually have?

These can be interpreted as probing the effects of realistic (expected) and possibly non-realistic (ideal) opinions on the outcome.

1.2 Target Trial

We design a target trial to clarify our causal questions, and to guide our methodology for feature selection and processing. We explore two different causal questions, resulting in two target trials that are identical except for their treatment strategies. These target trials are hypothetical and do not need to be ethical or feasible.

- Eligibility criteria: Individuals between the ages of 16 and 22. Selection of subject should result in a very geographically and socially sparse in order to avoid effects of subjects on each other.
- Treatment strategies:
  - For the first trial, at a predefined point in time, referred to as *time zero*, gather all participants and assign each of them, at that same moment, an **expectation regarding the**

[1] The code for this project is available in `https://github.com/itamar-github/causal_inference_project` .

number of children they expect to have in the future.[2] This number will be an integer randomly sampled i.i.d. from the distribution of the number of children observed in the population at that time.

  - The second trial will follow the same treatment procedure, but instead, we will assign each participant their **opinion on the ideal number of children in a family**.

- Follow-up period: gather data from the participants 40 years after *time zero*. Participants who have passed away or are unreachable for the follow-up will be excluded from the trial.
- Outcome: record the number of children each participant currently has.
- Analysis plan: analyze the effect of the treatment by calculating the average dose-response function. Additionally, Convert the treatment variable into a binary format using the rule $T_{\text{binary}} = \mathbb{I}[T > \theta]$, where θ is the median of the treatment. This allows us to compute the Average Treatment Effect (ATE) and estimate propensity scores.

1.3 Data

We use the National Longitudinal Surveys (NLS) [9], specifically the National Longitudinal Survey of Youth 1979 (NLSY79). NLSY79 is a series of surveys designed to collect information at multiple points in time on Americans born between 1957 and 1964. The survey includes 9,964 responders, who have been surveyed periodically from 1979 to 2020.

NLSY79 is very detailed and includes thousands of questions concerning various aspects such as beliefs, race, ethnicity, gender, family status, economic status, education, health, employment, and more. The main challenge of working with this dataset comes from its vast richness — selecting the appropriate variables to include in our model, without including unnecessary variables that might compromise our ability to measure the causal effect (for example, due to the "common support" condition).

Furthermore, due to the long time span of the study, there are many missing values due to responders that did not participate in all years the survey was conducted (for unknown reasons).

2 Feature Selection and Extraction

2.1 Background

Multiple works from different countries studied the factors affecting fertility [1, 6, 7, 10], and desired number of children [2, 8]. Unsurprisingly, the desired number of children is cited as a factor affecting fertility, and it was shown that the two are influenced by common factors. Using these sources, with additional features we found appropriate, we created the following list of variables we would have liked to have for our analysis.

- Age
- Economic status
- Educational status
- Ethnic group
- Gender
- Location
- Marriage status, age at marriage

[2] "Assignment" of such treatment is not currently possible, but similar concepts appear in science fiction.

- Mental health
- Nature of close relationships
- Number of siblings
- Number of children at treatment[3]
- Parents' features (marriage, living arrangements, employment, relationship with subject, etc.)
- Physical health
- Political affiliation
- Religion
- Sexual orientation

Some of these features are not present in the dataset, or are present but of low quality. Therefore, the final set of features we used is different. Before elaborating on the features in Section 2.4, we present the target features that we used.

2.2 Target

The target value — the total number of children the responders had, is not directly available in the dataset. Instead, throughout the years, the responders were asked for the number of biological children they had at the time of each survey. At first, it seemed like the latest possible instance of this question should be used as our target values, but after further investigation of the data we found two significant problems with this approach — (1) as the years progressed, the number of responders to the question declined substantially, and (2) this question was directed only towards females. We came up with two possible (partial) solutions.

1. The survey contains aggregate responses ("cross round"), i.e. responses not tied to a single year. One such question is the "Date of birth of n-th child". Summing all non-empty responses of a responder to this question is a lower bound to the actual number of children. This approach addresses both drawbacks presented earlier — (1) it applies to both males and females, and (2) it has no non-interview missing values. Unfortunately, the latter implies that the responses may not be accurate, as we know that there are many responders who did not participate in many of the surveys.
2. By combining the responses to the number of biological children at different years, we get a more accurate lower bound since we *retain the 'non-interview' missing values*. This still does not solve the problem that the question is exclusively addressed to females. We think that it is a reasonable concession, especially given the fact that sex would have been used as a feature in the analysis either way.

We ended up using the second option as our target variables. We identify two significant drawbacks of this approach — (1) it does not take into account adopted or step children. We think that the failure to account for step-children is somewhat mitigated by the fact that we only consider female responders. In addition, according to [5], adopted children comprised about 2 percent of U.S. children in 2007. While this number is not insignificant, we think that it is a reasonable source of error given the dataset. (2) The youngest responders to the survey were 14 in 1979, meaning that some of them were still in (late) reproductive ages at the time of their last survey (e.g. 41 at 2006). In light of 'age-at-last-birth' statistics (e.g. [3]), we find that this is reasonable, as most responders were older at the time of their last survey and considerably less likely to have had another child after their last survey.

[3]We study subjects who already had children at the time of the first survey separately.

2.3 Treatments

The choice of treatment values is straightforward — we used the responders' answers to the questions "what do you think is the ideal number of children for a family?", and "how many (more) children do you expect to have?", from the first survey (1979).

2.4 Features

The full list of features we used is shown in the attached jupyter notebooks. All features (other than the target values) were taken from the answers to the first survey (1979), at the same time that the treatment was observed.

We performed standard manipulations for all the features, such as transformation of categorical variables into one-hot (dummy) variables, and standardization of continuous variables.

We briefly mention the features which we manipulated in a nontrivial way:

- Age — since the age range in the data is large, and combines responders in different life stages, we divided the data into two groups — (1) ages 14-17, and (2) ages 18-22. The reason for choosing this particular division is that it was done in the survey in some questions, such as in the "what would a significant person think ..." questions presented later, which were only asked of 14-17 year old responders.
- "What would a significant person in your life think about if you told them that you decided (never to have children / to pursue a full time career and delay starting a family)?" — while the answer to these questions are ordered (1 strongly disapprove to - 4 strongly approve), they also contain an "don't know" option (8) and missing values. Hence, we decided to represent the responses as categorical values, so we can include a meaningful tokens for all possible values.
- # of children when first answered the survey (1979) — we split the data into two groups (1) responders who had already had children at the time of the first survey, and (2) responders who did not. We corrected for the number of children responders who had already had children at the time of the first survey when relevant.
- Sex — since our target observations are only based on female responders, we dropped the male responses.

We split the data into four groups — (1) young (age 14-17) with children (at 1979), (2) young without children, (3) mature (age 18-22) with children, and (4) mature without children. We analyzed each of groups (2), (3), and (4) separately, and did not use group (1) due to it being significantly smaller than the others (57 responses).

Some features that are not straight forward to interpret:

- SAMPLE ID - is a feature representing the group from which the sample was taken. The groups are determined based on sex, ethnicity (White, Black, Hispanic), and whether the sample was taken from the general population (cross-sectional), a supplemental group (oversampling civilian Hispanic or Latino, Black, and economically disadvantaged individuals), or the military.

We were unable to use some of the variables we mentioned in Section 2.1. Some were not used because they did not appear in the data, e.g. sexual orientation, mental health (at least at the time of the first survey). For others, we tried to use 'proxy' features when applicable, e.g. "health conditions that cause work limitations" for physical health.

In some cases, the data was of low quality so we decided to omit it, for example 'marital status of parents' wasn't available in the data, and we omitted 'proxy' features such as the answer to the question "mother and father living together" due to too many missing values.

2.5 Missing values

There are 5 kinds of missing values in the data:

- '-1' — refusal,
- '-2' — don't know,
- '-3' — invalid skip,
- '-4' — valid skip (e.g. answer to conditional question missing due to irrelevant condition),
- '-5' — non interview.

We treated 'missing value' in categorical variables as a category of its own. For continuous variables we used two approaches — (1) we dropped rows containing missing values when the total number of missing values in the column was small, and (2) when the number of missing values was large we replaced it with the median value.

3 Initial Statistical Analysis

There is a strong positive correlation of 0.79 (but not 1) between our two treatments: the number of children a candidate expects to have in the future (asked in 1979) and the candidate's opinion on the ideal number of children (also asked in 1979). Figure 1 shows the distributions of the two treatments across all three groups. The most common value for both treatments is 2 in all groups, and values greater than 4 are quite rare.

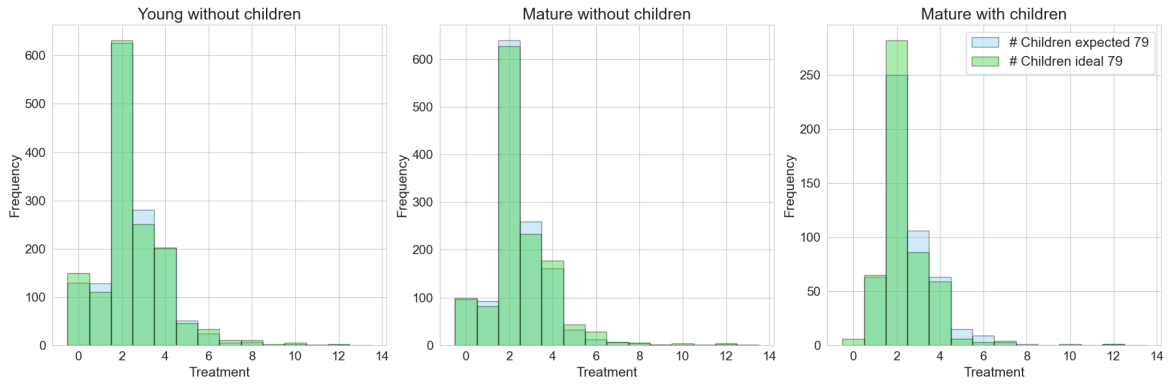


Figure 1: Histograms of the treatments 'expected # of children' and 'ideal # of children' in 1979, for each group of responders. Clearly, the treatments are very similar, yet not identical.

Figure 2 shows the distribution of the outcome (number of children participants had by 2016), which resembles the distribution of the treatments. Interestingly, in all groups, the most common value of 2 is less frequent in the outcome than in the treatment. In addition, in groups without children the frequencies of 0 and 1 are much higher in the outcome than in the treatment. This suggests that many candidates end up having fewer children than they expected or considered ideal before becoming parents.

We looked at the correlations between the features and the outcome in the different groups. The treatments "TOT # CHILDREN EXPCT HAVE 79" and "# CHILDREN IDEAL FOR FAMILY 79" are positively correlated with the outcome ($\rho \approx 0.1$ for responders without children, and $\rho \approx 0.2$ for responders with children at the time of the first survey), with the former showing a stronger correlation. In Figure 3, we plot the top 4 largest positive and negative correlations of features with

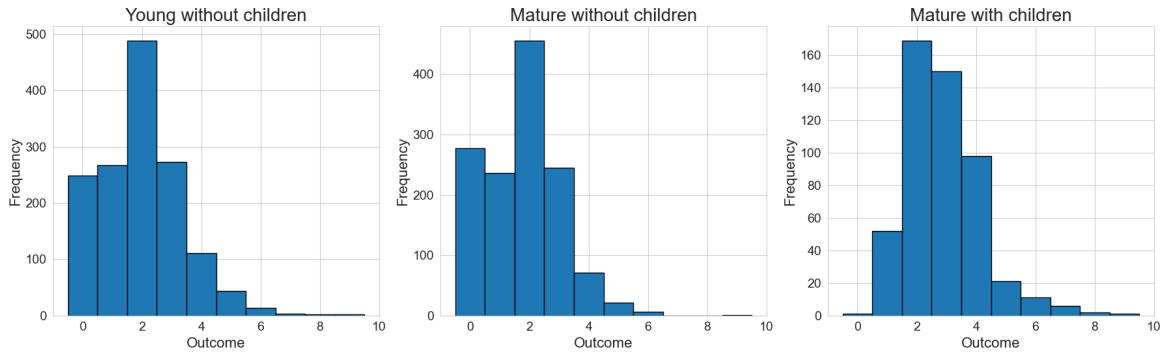


Figure 2: Histograms of the outcome's distribution for each group of participants.

the outcome, for each group. Features with strong correlations include the number of siblings, poverty status, and ethnicity (e.g., being Black, Hispanic, Mexican). Interestingly, for the group of mature participants who already had children in 1979, educational status features, such as completing less than 12 grades or not being enrolled in college, are also highly correlated.

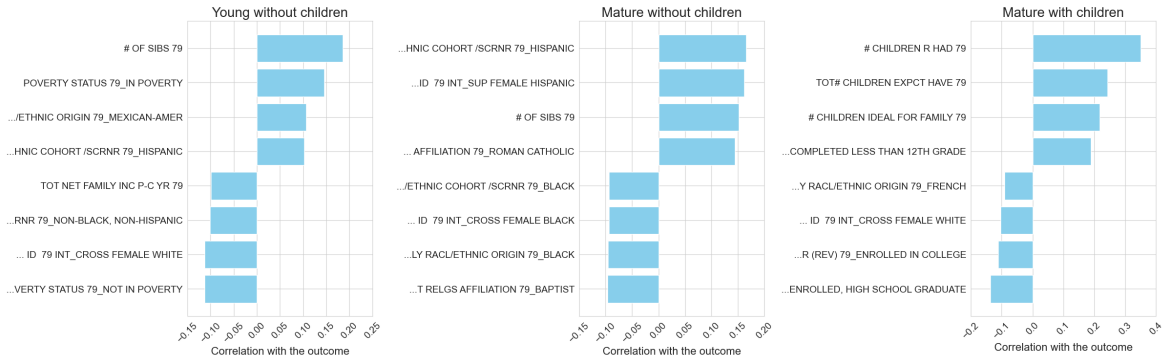


Figure 3: Top 4 largest positive and negative correlations with the outcome for each group.

4 Identification Assumptions

In this section, we discuss the identification assumptions that enable the estimation of the causal effect of the treatments using the data.

Stable Unit Treatment Value Assumptions (SUTVA).

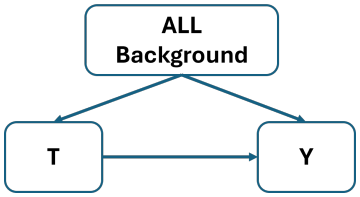
1. "The potential outcomes for any unit do not vary with the treatments assigned to other units" — in our case the dataset contains many related responders that can affect each other. In order to identify such relationships we used the 'household ID' feature, that uniquely identifies responders from the same household. After removing male responders and cleaning the data we were left with 3958 total responders, with 3337 unique households. In order to comply with this assumptions we kept only a single responder from each household.
2. "For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes" — in our case, treatments for all subjects were the same. Our

analysis is divided into two parts. In Section 5.1 we use a binary treatment version (above or below median) that does not satisfy this assumption. We acknowledge that this is problematic, yet find that it is reasonable enough for our purposes as it is only one part of the analysis. In Section 5.2 we analyze the original treatment values directly. In this setup the assumption is satisfied.

Consistency. "For a unit that receives treatment T, we observe the corresponding potential outcome Y_t" — as in all data based on self report, there may be some noise in the recorded values (this is true for the entire dataset, not just the treatments). In this case, we think that the problem is not severe due to the fact that (1) the data is anonymous, and (2) responders were given the option not to answer the question (see Section 2.5). Nonetheless, the data was collected through interviews, and although family members or other acquaintances were not part of the interview, they were possibly nearby, which might have led to some respondents providing false answers.

Ignorability. "No unmeasured confounders". We selected a wide range of background covariates based on our intuition and previous work on the subject (see Section 2.1). In addition, in the groups of responders who did not have any children at the time of the first survey, there is a clear temporal relationship between the outcomes and the rest of the features and treatments, so we find it unlikely to have any 'collider' structures in a causal graph of the model (see Figure 4). For this reason we also did not include the answers to the question "how many children do you want to have?", which would have had a complex causal relation with the treatments and outcomes.

Figure 4: **Causal graph for groups without children.** We believe that we have captured a significant part of the background covariates.



The case of the group that already had children at the time of the first survey is more complex, as there is a nontrivial relationship between the treatments, outcome, and the feature "current number of children" (recorded in 1979). We are therefore less confident in the degree to which ignorability is satisfied in this group.

Common support. "Which means $Pr[T = t|X = x] > 0, \forall t, x$". In each group separately, we trained multiple types of models[4] to predict the binary treatment values from the other features (excluding the other treatment type and the outcome). Each trained model was then calibrated on a different calibration set. While some models trained to perfect interpolation of the training set, the best F_1 5-fold CV score (≈ 0.4 for all groups) was achieved by basic models such as naïve Bayes and logistic regression (e.g. Figure 5). This suggests that the common support assumption holds.

5 Results

We generated two types of results in our analysis: one by converting the treatment variable into binary values, allowing us to compute the Average Treatment Effect (ATE) and propensity scores (Section 5.1); and another using the original treatment, which is a continuous variable, enabling us to plot the average dose-response function (Section 5.2).

[4]Gradient boosting, KNN, logistic regression (L2, elastic), naïve Bayes, random forest, SVM (linear, RBF), and XGBoost. Hyperparameters were tuned using 5-fold CV. For full detail see our implementation.

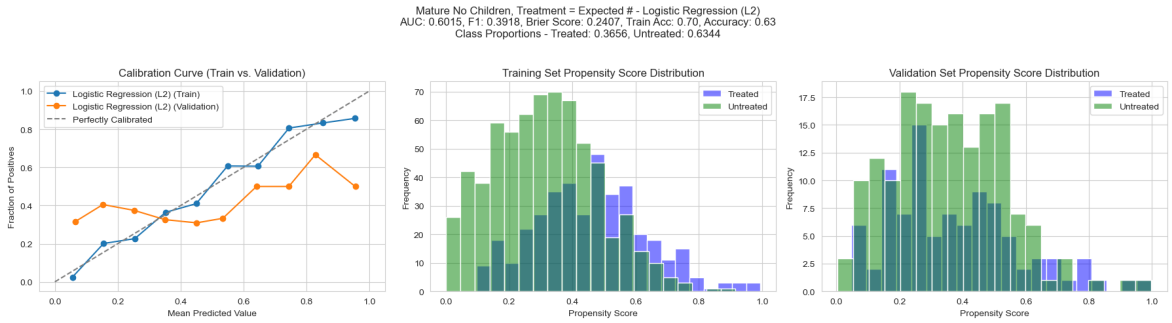


Figure 5: **Propensity score estimation** of the treatment "expected # of children > 2 (median)" for mature responders (18-22) without children at first survey. Behavior in other groups is similar.

5.1 Analysis with Binary Treatments

Both treatments, T_1 (expected number of children) and T_2 (ideal number of children), were converted into a binary variable using the threshold $T_{i_{\text{binary}}} = \mathbb{I}[T_i > 2]$, $i = 1, 2$. We chose a threshold of 2, as the median value of both T_1 and T_2 is 2; also, the mean values are 2.47 and 2.50 respectively. The binary variable $T_{1_{\text{binary}}}$ can be interpreted as follows: does the participant expect to have more children than the median value in the population? And similar for $T_{2_{\text{binary}}}$.

5.1.1 Methods

The following methods were used in order to estimate the Average Treatment Effect (ATE), where $\text{ATE} = \mathbb{E}[Y_1 - Y_0]$.

- S-learner: fit a model f with T as feature on the entire dataset. And then compute $\text{ATE} = \frac{1}{n}\sum_i f(x_i, t = 1) - f(x_i, t = 0)$. We used three models to compute the ATE: Support Vector Regression (SVR), Random Forest Regressor, and Gradient Boosting Regressor. The final ATE was the average of their individual ATEs. These models were chosen because they achieved the lowest MSE on a test set we created for predicting Y.
- T-learner: fit two separate models f_0 and f_1 on treated and control samples. And then compute $\text{ATE} = \frac{1}{n}\sum_i f_1(x_i) - f_0(x_i)$. We used the same models as in the S-learner.
- Matching: For each unit i with treatment t_i, find the k closest datapoints $j(i)$ with treatment $1 - t_i$, and use it to compute the CATE for this sample. For example, in case $t_i = 1$: $\text{CATE} = y_i - \frac{1}{k}\sum_{l \in j(i)} y_l$. The ATE is the average of the CATE values. We used $k = 9$ because it achieved the lowest MSE on a test set we created for predicting Y.
- IPW: based on the analysis from Section 4 we selected an SVM model as propensity score estimator in all groups. We chose this model because it had comparable predictive power to other models, yet unlike the others, its predictions were very conservative (far from 0 and 1; see Figure 6) thus reducing variance and somewhat accounting for the low predictive performance.

5.1.2 Results and Discussion

The results are summarized in Figure 7.

Discussion. As can be seen, the ATEs are significantly positive, except for the IPW estimates (based on bootstrap estimation) across all groups, suggesting that expecting a large number of children ($T_{1_{\text{binary}}}$) or considering a large number as ideal ($T_{2_{\text{binary}}}$) cause candidates to have more children in the future. The ATE is largest in the "mature with children" group. We hypothesize that this

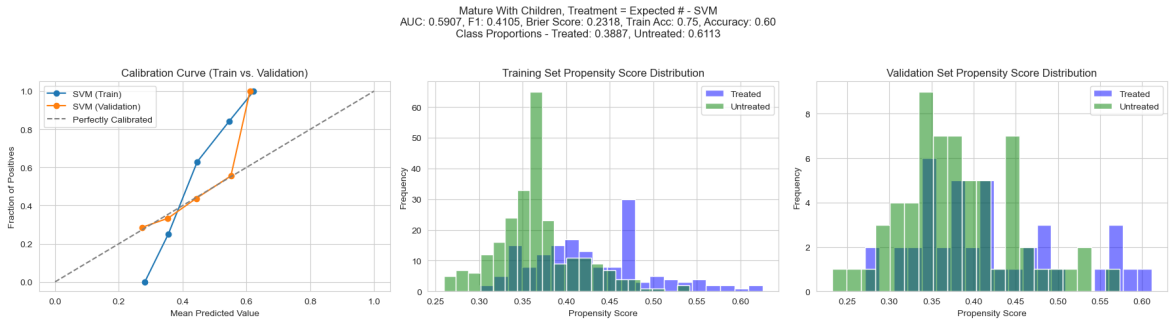


Figure 6: **Propensity score estimation** with RBF-SVM + Platt scaling, of the treatment "expected # of children > 2 (median)" for mature responders (18-22) with children at first survey. The predicted propensity scores are (empirically) bounded away from either 0 or 1. This is a desired property given the poor predictive performance of the models.

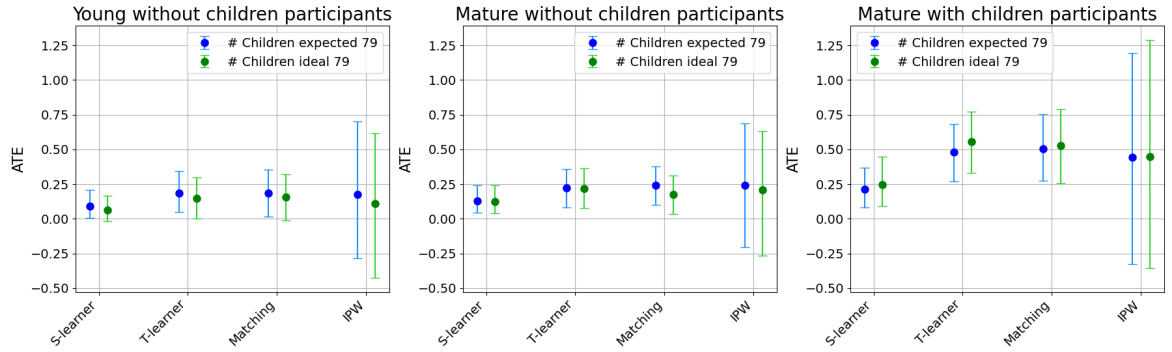


Figure 7: **Estimated average treatment effects (ATE) per group**. Confidence intervals (95%) were calculated with 1000 bootstrap samples.

is because more mature or ready participants have more realistic and well-considered expectations and opinions regarding the number of children they will have in the future. Consequently, their expectations and opinions have a stronger effect on the actual number of children they eventually have. These results suggest that it is interesting to consider the relationship between our treatment variables and wether responders had already had children when they answered the survey, but we did not expand on it in this work. It also appears that $T_{1_{\text{binary}}}$ has slightly more influence on the outcome in the groups of candidates without children. However, this trend reverses in the "mature with children" group, where $T_{2_{\text{binary}}}$ has a slightly greater effect.

Weaknesses. Estimating the ATE is challenging because we lack a method to evaluate the performance of our models, unlike in standard ('predictive') machine learning where a test set can be used. As a result, although our different methods produce varying ATE estimates (with the S-learner providing smaller estimates), we cannot determine which model yields the most accurate results. Additionally, the confidence intervals are quite large, particularly in the "mature with children" group. Despite this, the results seem to indicated a positive effect of the treatments on the outcome.

5.2 Analysis with Continuous Treatments

5.2.1 Methods

Dealing with a continuous treatment requires a different formulation and tools than the ones we learned in the course.

Setting. For each sample i and treatment value $t \in \mathbb{T}$, we define a set of potential outcomes $Y_i(t)$, also referred to as the unit-level dose-response function. We are interested in estimating the average dose-response function $\mu(t) = \mathbb{E}[Y_i(t)]$ for every $t \in \mathbb{T}$. This will give us insight into how the continuous treatment affects the behavior of the outcome. As shown in [4], under certain assumptions (a continuous generalization of the ones we saw in class), it is possible to estimate $\mu(t)$ from data. Most of these assumptions are identical or very similar to the district-specific ones we discussed in Section 4.

Method. Estimating $\mu(t)$ requires a few steps:

- Estimate the Generalized Propensity Score (GPS) $r(t, x) = f_{T|X}(t|x)$.
- Estimate the conditional expectation of the outcome as a function of the treatment and the GPS $\beta(t, r) = \mathbb{E}[Y|T = t, R = r]$.
- Average $\beta(t, r)$ over the GPS for each $t \in \mathbb{T}$, giving an estimate of the dose–response function $\hat{\mu}(t) = \mathbb{E}[\beta(t, r(t, X))]$.

Similar to the propensity scores used in the binary case, the GPS allows us to compensate for biases present in our data. We will estimate of the dose–response function using the implementation from [4], which also performs bootstrap resampling to obtain confidence intervals.

5.2.2 Results and Discussion

We estimated the average dose-response function for each of our participant groups, both for T_1 and T_2, as shown in Figure 8.

Discussion. Our results are consistent with the ATEs we computed using the binary treatments (Section 5.1). Specifically, we observe that the outcome tends to increase with the increment of both treatments for small treatment values, before approximately stabilizing with growing confidence margins for large treatment values. This increase is most significant in the "mature with children" group. Moreover, the dose–response function estimations for both treatments exhibit almost the same behavior across all groups. Therefore, we cannot conclude that one treatment has a greater effect than the other.

Weaknesses. To estimate the dose-response for specific treatment values, we use only a small subset of the data corresponding to those values, which may result in various inaccuracies. The confidence intervals are widest at the smallest and largest treatment values due to less available data (see Figure 1). We also observe a decrease in the dose-response for the largest values of T_1 and T_2, yet we cannot conclude that this is indeed the case as it is accompanied by an increase in the confidence intervals. This phenomenon, which also been observed in [4], likely also stems from the small amount of data. Additionally, the optimization did not fully converge for some T_1 and T_2 values due to the extremely low number of samples (this happened for large treatment values, and determined the upper edge of the treatment axis in Figure 8).

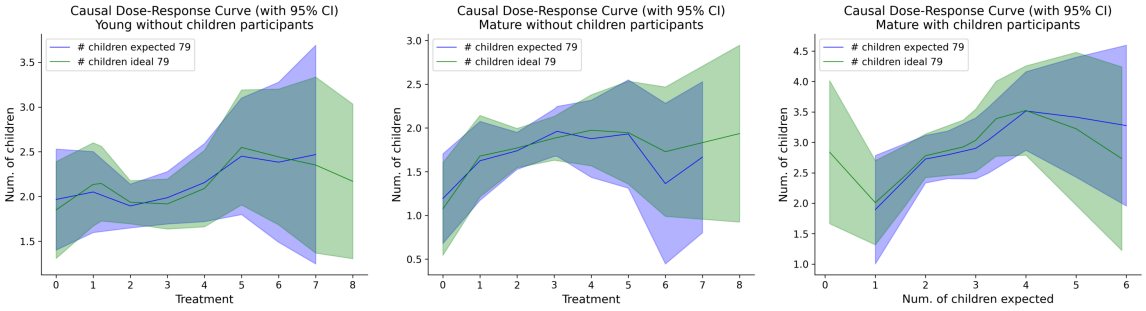


Figure 8: **Average dose–response** function for each of our participant groups, with 95% confidence intervals.

6 Conclusion

In this project, we explored the following causal questions: how does a person's expectations, and opinions about the ideal number of children in a family, influence the actual number of children they eventually have. We pre-processed and analyzed features from the NLSY79 [9] dataset, focusing on three distinct groups: young responders without children at the time of the first survey, mature responders without children, and mature responders with children. We applied multiple causal inference methods to estimate the effects of the treatments from both binary and continuous perspectives. Our results align with our expectations, indicating that expecting a large number of children (T_1) or considering a large number as ideal (T_2) influences candidates to have more children in the future. These effects appear strongest in the "mature with children" group, possibly because their expectations and opinions about children are more realistic and well-considered. However, from our current results, it is challenging to determine if one treatment has a greater effect than the other.

References

[1] R. Adhikari. Demographic, socio-economic, and cultural factors affecting fertility differentials in nepal. *BMC Pregnancy and Childbirth*, 10(1):19, Apr. 2010.

[2] A. B. Bakilana and R. Hasan. The complex factors involved in family fertility decisions, May 2016.

[3] L. Carlson and K. B. Guzzo. Median age at last birth. *Family Profiles, FP-21*, 5, 2021.

[4] R. W. Kobrosly. Causal-curve: a python causal inference package to estimate causal dose-response curves. *Journal of Open Source Software*, 5(52):2523, 2020.

[5] Office of the Assistant Secretary for Planning and Evaluation. National Survey of Adoptive Parents (NSAP) — aspe.hhs.gov. `https://aspe.hhs.gov/national-survey-adoptive-parents-nsap`.

[6] M. Ranjbar, M. K. Rahimi, E. Heidari, S. Bahariniya, M. Alimondegari, M. H. Lotfi, and T. Shafaghat. What factors influence couples' decisions to have children? evidence from a systematic scoping review. *BMC Pregnancy Childbirth*, 24(1):223, Mar. 2024.

[7] T. J. Samuel. Social factors affecting fertility in india. *Eugen Rev*, 57(1):5–15, Mar. 1965.

[8] K. Sarvestani, S. Khoo, N. Malek, S. Mat Yasin, and A. Ahmadi. Factors influencing the desired number of children among married women in the reproductive age and its implications for policy making. *Women's Health Bulletin*, Inpress, 08 2016.

[9] U.S. Bureau of Labor Statistics. National longitudinal surveys, 1979.

[10] Z. Xiang, X. Zhang, Y. Li, J. Li, Y. Wang, Y. Wang, W.-K. Ming, X. Sun, B. Jiang, G. Zhai, Y. Wu, and J. Wu. Fertility intention and its affecting factors in china: A national cross-sectional survey. *Heliyon*, 9(2):e13445, 2023.