Evaluating Effective Interventions on Individuals in Infelicitous Eventualities

An initial proposal report discussing a pair of Causal Inference tasks
and a proposed approach for solving these two problems.

2100816

February 24, 2022

Registration number: 2100816
Project: Causal inference
Link to GitHub: https://github.com/11BelowStudio/ce888

Executive summary (max. 250 words)	96
Introduction (max. 600 words)	599
Data (max. 500 words/dataset)	*(538 + 459)* = 997
Methodology (max. 600 words)	635
Conclusions (max. 500 words)	473
Total word count	2800

Table 1: Word counts for each section.

Contents

Abstract

This document is the formal proposal document for a causal inference investigation into the the IHDP[1] [2] and JOBS[3][4][5] datasets.

This document introduces the context for the problem, discusses some findings from an initial exploration of these datasets (providing visualizations of these datasets to supplement this), and provides a proposed methodology for how the next stages of this investigation shall be achieved.

The preliminary investigation has identified some potential roadblocks for the latter parts of the investigation, which could pose a few barriers to the potential for meaningful conclusions to be reached, however, these are not insurmountable.

1 Introduction

This project involves two datasets: the Infant Health and Development Program (`IHDP`)[1][2] and `JOBS`[3][4][5].

Both of these datasets contain information about individuals (`x`), whether or not the individuals received some 'treatment' (`t`), and a `y` outcome for the individuals. The task I have been given for these datasets is to find the causal relationships within these datasets, to assess whether or not the treatments (`t`) given to the individuals have had any effect on the outcomes (`y`).

As discussed by Hill and Stuart, 'Causal Inference' is the term used to refer to the overall task of investigating how a 'causal variable' may influence an 'outcome', and what conclusions can be drawn from that. There is a particular interest in trying to predict 'the outcomes that could manifest given exposure to each of a set of treatment conditions', allowing one to perform 'comparisons between these 'potential outcomes"[6]. This act has practical applications that serve genuine benefits besides the existential flex of predicting an alternative timeline, for example, Glass et al mention how this act of identifying causal relationships has formed the backbone of public health policy and modern medical practice, and emphasize the importance of using causal inference to establish the effects of interventions, not just underlying causes, to allow meaningful interventions to be made as and when necessary[7]. This relates to the concept of Causal Decision-Making, which, by itself, does not need the counterfactuals and causal effects to be calculated (only needing an estimator of `y` given `x` and `t`). Fernández-Loría and Provost do stress the importance of not overcomplicating that particular task through unnecessarily introducing counterfactuals, however, they do point out that causal inference is vital for evaluating the success of these causal decisions[8].

In context of the `IHDP` and `JOBS` datasets; `IHDP` concerns the cognitive development of prematurely-born children, with an intervention in the form of additional support being given to the families in the control group, with the intent being to, as the name implies, support the health and development of these infants, throughout their childhoods[1][2]. This intervention could provide many benefits to the parents and the child besides being able to score high on cognitive ability tests (the subset of the `IHDP` dataset I have access to for this project only has a cognitive ability test result score as a `y` outcome though), however, if a meaningful result for the intervention cannot be demonstrated, it is unlikely that resources would be allocated to allow this intervention to be sustained long-term, for more individuals. Furthermore, if it is found to be ineffective, that could be seen as a motivation to find other potential interventions that may turn out to be more effective (and worthy of being used in the long term).

`JOBS`, on the other hand, concerns the effect of a support program on helping individuals who are unemployed to gain employment (with a rather large control group consisting of individuals who were not on this support program)[3][4][5]. This dataset arguably does have a somewhat uninformative `y` value, just like `IHDP`, as this `y` is merely a binary value (employed or unemployed), regardless of the variety of employment (whether that 'employment' be in the form of a stable, well-paid job, or underemployed without any job security). However, just like in the situation of `IHDP`, if the treatment doesn't have any positive effect on whether or not an individual can successfully gain employment (especially when considering the employment outcomes for other, similar, individuals who were not receiving this additional support), this particular support, being unfit for purpose, would need to be replaced by a more effective intervention.

2 Data

2.1 IHDP - The Infant Health Development Program

IHDP consists of real data regarding the effects of providing additional support to the families who had premature babies on the development of aforementioned babies, measured through the means of IQ tests[1][2]. The version of this dataset used in this task contains factual and (simulated) counterfactual data; however, I shall only use the counterfactual data for evaluating fully trained models.

This dataset concerns a population of 750 individuals, all of whom have 25 x values, a t value indicating whether they were in the treatment/control group, as well as a factual yf value (indicating the true y value recorded from that individual during the experiment), and a counterfactual ycf value (derived via a simulation, simulating the outcome for that individual if they were in the opposite treatment/control group to the group they actually were in). There is also an ite value, providing the individual treatment effect for the individuals. The y values are continuous values.

A visualization of this dataset can be seen in 4 (along with an overview of the factual data within 5, and of the counterfactual data within 6).

Of the 25 'x' values, 6 of them are in a continuous range, whilst the other 19 are binary. One of these non-binary 'x' values, x3 , appears to be limited to one of four discrete values; however, as I have been unable to conclusively find out what x3 truly means, and as that means it could possibly have a different value, I shall not normalize that into discrete values, to allow leeway for potential future data, which may have a different value.

Figure 1 indicates a lack of much correlation in this dataset, barring a few notably higher correlations.

There is a rather strong negative correlation between x5 and ite (and, in extension, x5's relation to t0 and t1), to see if this correlation truly indicates whether a higher x5 can cause the treatment to have an unintended effect. This correlation is also illustrated in Figure 4, showing this strong downward trend. However, looking at the x5 data in Figure 5, we can see some individuals with high x5 values in the untreated group achieved rather high yf scores (but only with a x5 and yf correlation of 0.36), so one could argue that the negative ITE correlation could be due to the counterfactual simulation not allowing similar ycf values to be reached (with counterfactuals illustrated in Figure 6).

Other notable correlations include a strong negative correlation of -0.87 between x3 and x13, indicating a potential causal link. A weaker positive correlation (0.68) is present between x0 and x1 (with slightly weaker negative correlations between those and x2). These correlations could be a sign of causation between these variables, or they could be an indicator of an external confounder, or perhaps neither, which justifies some further investigation.

2.1.1 The Causal Questions

1. To what extent does each x predict y?
2. To what extent does t predict y?
3. To what extent do the x values predict each other?
4. To what extent do the x values predict the ite values?

2.1.2 Metrics to use

As this dataset contains full counterfactual/ITE data, I am able to use Precision in Estimation of Heterogenous Effect ($\epsilon PEHE$) and Average Treatment Effect (ϵATE) in order to measure the correctness of the learner's estimations of individual treatment effects[9].

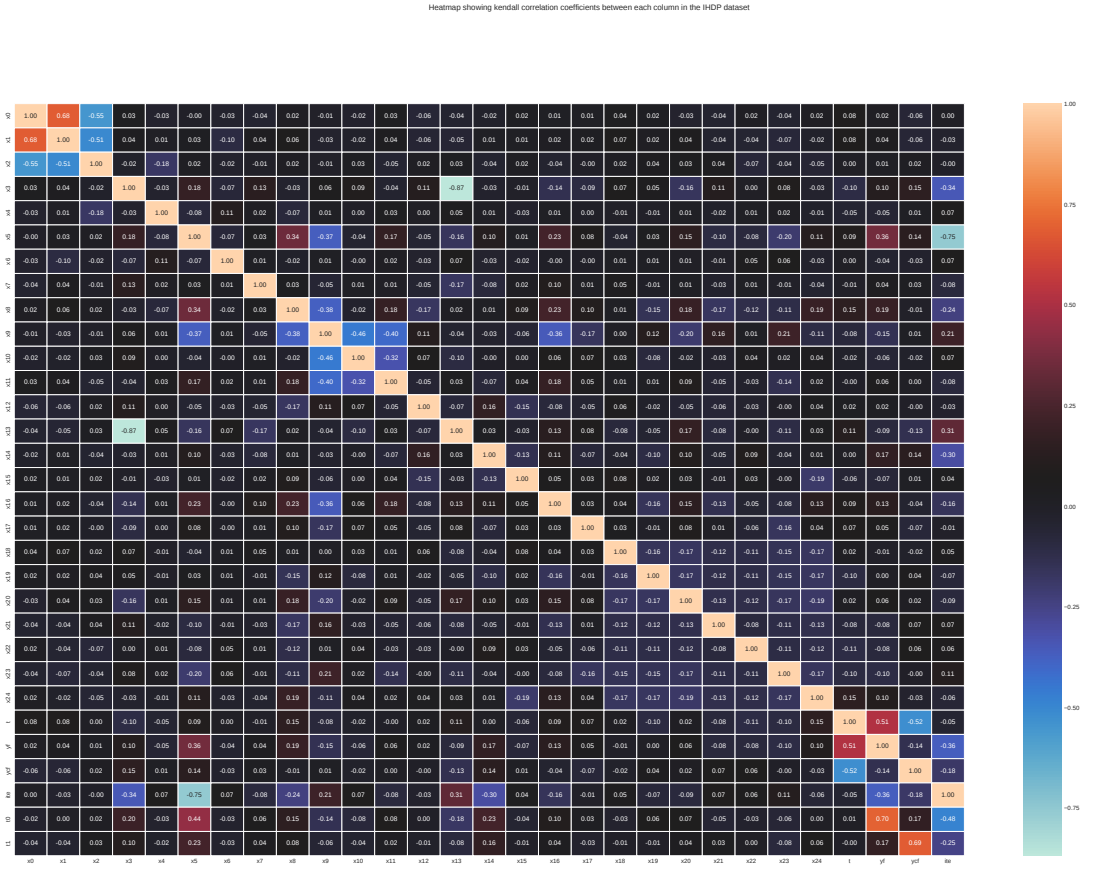

Figure 1: Correlation heatmap for the IHDP dataset
This annotated heatmap shows the correlation coefficients for the values in each column of the IHDP dataset (along with additional `t0` and `t1` columns). `t0` contains the `y` values for each individual for the case where `t=0`, and vice versa for `t1` (with these columns being included for ease of looking at overall treatment/control outcomes).
There isn't much clear correlation between the values of each column in this dataset, barring a few outliers. There is a rather strong positive correlation between `t` and `yf` (with a slightly stronger negative `t`/`ycf` correlation), indicating that being in the treatment group correlates to a higher `y`, but whether or not this is due to causation does need to be investigated further.

2.2 JOBS

`JOBS` contains data about 3212 jobseekers (`x`), whether or not they received support to get a job (`t`), and whether or not they were successful in gaining employment (`y`, with a value of 0 or 1)[3][4][5]. There is no counterfactual data within this dataset, and the dataset is derived from experimental and observational data (with this being denoted by `e`). This dataset is incredibly unbalanced, with only 297 individuals being in the treatment group, and only 482 individuals with `y=0`, which will pose some problems with regards to avoiding overfitting.

Furthermore, many of the `x` values are binary, with the non-binary `x` values having particularly extreme ranges (requiring the use of a log scale to graph those appropriately, as can be seen in Figure 7), heavily necessitating some normalization before training is attempted.

Looking at the correlation heatmap for `JOBS` in figure 2, we can see that there is significant correlation within the dataset, with perfect positive correlation between `x0`,`x8`, and `x9`. `x1` and `x10` also have perfect positive correlation, along with `x11` and `x6`, and again with `x12` and `x7` (with many other sets of variables having imperfect yet rather high magnitudes of correlation). Upon looking at figure 7, these identified areas of perfect correlation aren't particularly unbalanced, but all of them are `x` values where `x` is in a range, instead of being binary. Interestingly, there isn't any strong correlation between `y` and any other variable, whilst `t` and `y` have a weak negative correlation of -0.07, suggesting that the treatment doesn't

Figure 2: Correlation heatmap for the JOBS dataset

have any bearing on the overall outcome. This implies the presence of some confounder having a stronger bearing on the `y` value, or that, if there is a causal relation between `x` and `y`, a combination of `x` values is what indicates the outcome.

2.2.1 The Causal Questions

1. Does `t` have any meaningful effect on `y`?
2. Do any `x` values meaningfully predict `y`? Do `t` or the other `x` values have an impact on this?
3. Do any `x` values have a causal interrelationship?
4. How can we measure the individual treatment effects due to a lack of counterfactuals?

2.2.2 Metrics to use

Due to the binary nature of `y`, it is possible to approach some aspects of this dataset from the perspective of a classification rather than a regression approach, meaning that we could use classification-based metrics, such as logistical loss to estimate the accuracy of `y` estimators. [10]

For measuring any Conditional Average Treatment Effects, the lack of counterfactuals requires either a Policy Loss metric (as explained in [11]), or a 'Average Treatment Effect on treated' metric[9]. However, as there are very few individuals who were treated, attempting to use the latter to assess performance on the full dataset is very likely to result in it being overfit. Therefore, Policy Loss shall have to be used instead.

3 Methodology

3.1 Preparing the learners

Before starting any training of the machine learners, I shall split the datasets into a learning set (to use in training via K-fold cross validation) and a validation set.

Furthermore, for sake of consistency, I shall use an RNG instance with a seed of 42 for the notebook, allowing reproducible results.

I shall use all of the counterfactual data of `IHDP`, along with 10% of the factual data (aiming to stratify based on treatment/control and the ITE counterfactual data) as its validation set, leaving the remaining factual data available for k-fold cross validation training.

For the JOBS dataset, I shall use 20% of the full data as the validation set (20% of treatment/control, aiming to stratify based on the outcomes), leaving the remaining 80% for training/testing.

When performing the testing, I shall be using scikit-learn's pipeline API, using pipelines with the structure explained in 3, for ease of re-running experiments, and using `GridSearchCV` to perform hyper-parameter tuning during K-fold training.

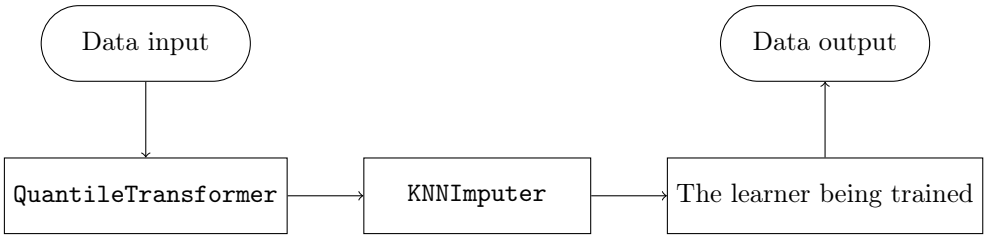

Figure 3: Structure for the pipelines to be used for training

3.2 Simple Learners

To obtain feature importances, I will attempt training a `RandomForestRegressor`, an `ARDRegressor`, and a couple of `AdaBoostRegressors` (consisting of the aforementioned other regressors) to predict the `y` outcomes for individuals given their `x` attributes and whether or not they have been treated (using `GridSearchCV` to handle hyper-parameter optimization). The `RandomForestRegressor` shall be used due to how it conveniently provides a method for returning feature importances @[12]. Obtaining these importances is slightly more convoluted with the `ARDRegressor`, however, there is a workaround to obtaining these. Despite the necessity of this workaround, the `ARDRegressor`, being a Relevance Vector Machine, aims to perform regression by trying to weight the inputs based on evidence maximization, in turn indirectly providing feature importances, without relying as much on a random seed. [13][14] The `AdaBoostRegressor` is an extension of these; effectively applying multiple instances of the same regressor to the problem, weighting each one appropriately to catch the cases which were missed by the previous regressors within the ensemble. [15]

The classification accuracy for this stage shall be assessed via `r2` scores, comparing the predicted `y` for the individual (from their `x` and `t`) to the factual `y` value. `r2` has been chosen due to how it provides a clear indicator of how good the estimator is compared to returning the expected `y` value (on top of how close the estimator is to returning 'true' `y` values), clearly indicating how good the predictions are. [16]

3.3 Propensity Score re-weighting

Similar to the 'Simple Learners' stage, I shall try using the identified classification models to predict propensity scores for each dataset (by aiming to predict `t` given `x`), and then using the identified 'best' classifier to predict what class they will be in (and giving that to the Inverse Propensity Score function to produce the actual sample weights). Unfortunately, due to the `ARDRegressor` not supporting sample weights, I will need to replace it with a standard `BayesianRidge` regressor for this, but it has a similar end result [17].

I will then repeat the process outlined in 'Simple Learners' with these weighted models, and compare the results of the learners with weighted samples to the performance of the unweighted learners.

3.4 Advanced CATE estimators

I shall attempt to compare the performances of the `XLearner`[18], `CausalForestDML`[19], and `ForestDRLearner`[20] CATE models, implemented as part of EconML[21], giving them the 'best' trained propensity and y-prediction models, and evaluating their performances on predicting treatment effects.

Each of these models come from different categories of CATE model, thereby allowing different aspects of the conditional treatment effect to be modelled.

During training, I shall evaluate the performances of these models on their 'policy risk' scores of the factual dataset (using the equation specified in the assignment brief) [11]. After training is done, I can evaluate the overall performance of the `IHDP` model via Precision in Estimated Heterogenous Effects, due to the presence of counterfactuals (which will not be possible for `JOBS`)[9].

4 Conclusions

This project seems somewhat feasible. The `IHDP` dataset looks like it is less likely to cause problems later on, as, due to it containing data for individual treatment effects and counterfactuals, it doesn't suffer significantly from an imbalance between treatment and control groups (although the factual data, which I need to use for testing, does), and this does permit some slightly easier evaluation of accuracy predictions after training is complete. However, from a more cynical perspective, that arguably does prompt the question of whether or not the findings I can derive from this will be of much practical use, as the presence of counterfactuals suggests that someone else already has fully analysed this dataset to the point of being able to fully simulate it, so, if a client were to realistically request an analysis of this particular dataset, one could, in theory, simply refer to existing literature, as this dataset isn't devoid of prior analysis.

However, the JOBS dataset could pose some significant problems. Besides the lack of counterfactuals, the massive imbalance between the sizes of the treatment and control groups compounds the existing imbalances between the quantities of individuals with each outcome (and for each `x` value). Of course, estimating the effect of a treatment (and what the effect would have been without the treatment) is one of the key tasks in causal inference, and it's generally physically impossible to get counterfactual data without having already analysed the factual data enough to accurately simulate the counterfactuals (unless, of course, one manages to somehow prove the many-worlds theorem and find the correct other world with the perfect counterfactuals available, but, if the necessary preconditions for that were to be met, this current task would probably be the least of one's concerns), so the lack of counterfactuals in `JOBS` is understandable. However, the limited treatment group data (presumably due to the criteria which individuals had to meet in order to be allowed into the treatment group, as explained in [5]) is likely to pose problems, especially if the predictor is provided with unseen data for an individual who would not have been included in the treatment group, and is expected to predict what the outcome would have been if they had received treatment (due to a lack of similar individuals in the treatment group to compare that individual to).

To conclude, I do not anticipate that any truly meaningful conclusions will be reached from my analysis of these datasets, but it is by no means impossible for that to happen, assuming that I am actually approaching this task from the correct angle. That said, I do not see merit in promising things which I know I cannot guarantee, so, whilst I cannot formally promise the meaningful outcome, I can at least promise to attempt delivering such an outcome within any future analysis of these datasets.

A Appendix: The visualizations of the datasets

These figures are located in this appendix simply because they're too large for LaTeXto neatly insert them within the main content of this report in a location that makes sense.

A.1 IHDP dataset visualizations

Here are the graphs for the IHDP dataset.

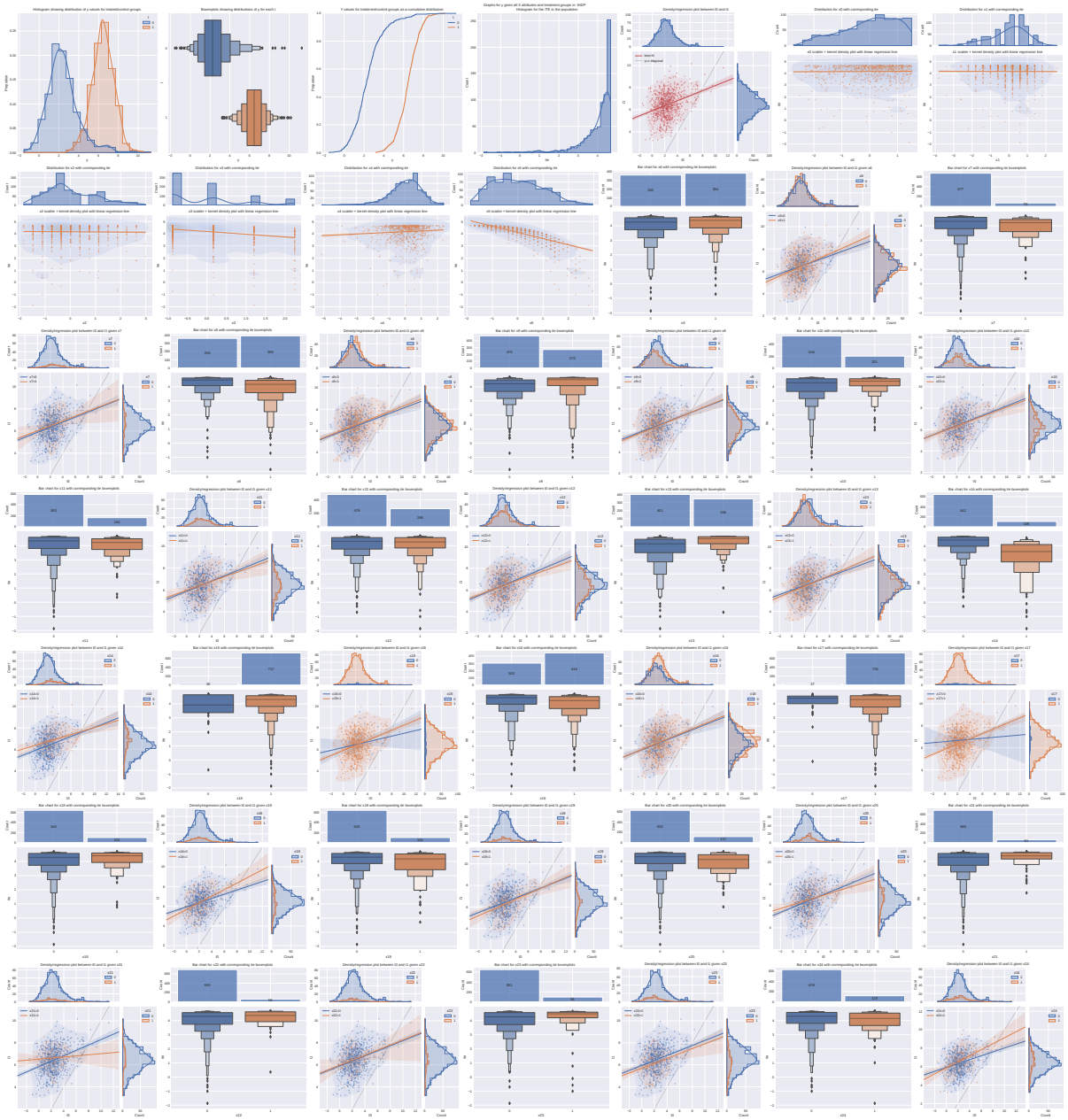

Figure 4: Several graphs for the IHDP dataset, including counterfactuals

These graphs show how the `ite` values (along with `t0` and `t1`) vary for each individual, based on the value of each `x` variable for each individual. `t0`, `t1`, and `ite` are based on known counterfactual data (end result being that `t0` contains the `y` value for the case where the individual was in the control group, and vice versa for `t1`).

Looking at the density/regression plot between `t0` and `t1`, we can see a general improvement in the `y` scores for the population when receiving the treatment, with few individuals falling below the `y=x diagonal` line (these individuals being those who had a negative `ite`).

We can see a somewhat clear negative correlation between `x5` and `ite`, and we can also see that, for

several of the binary-valued `x` values, there is a rather large imbalance in the quantities of individuals in the dataset who have each value, which could limit the amount of useful information we could potentially gain from these variables.

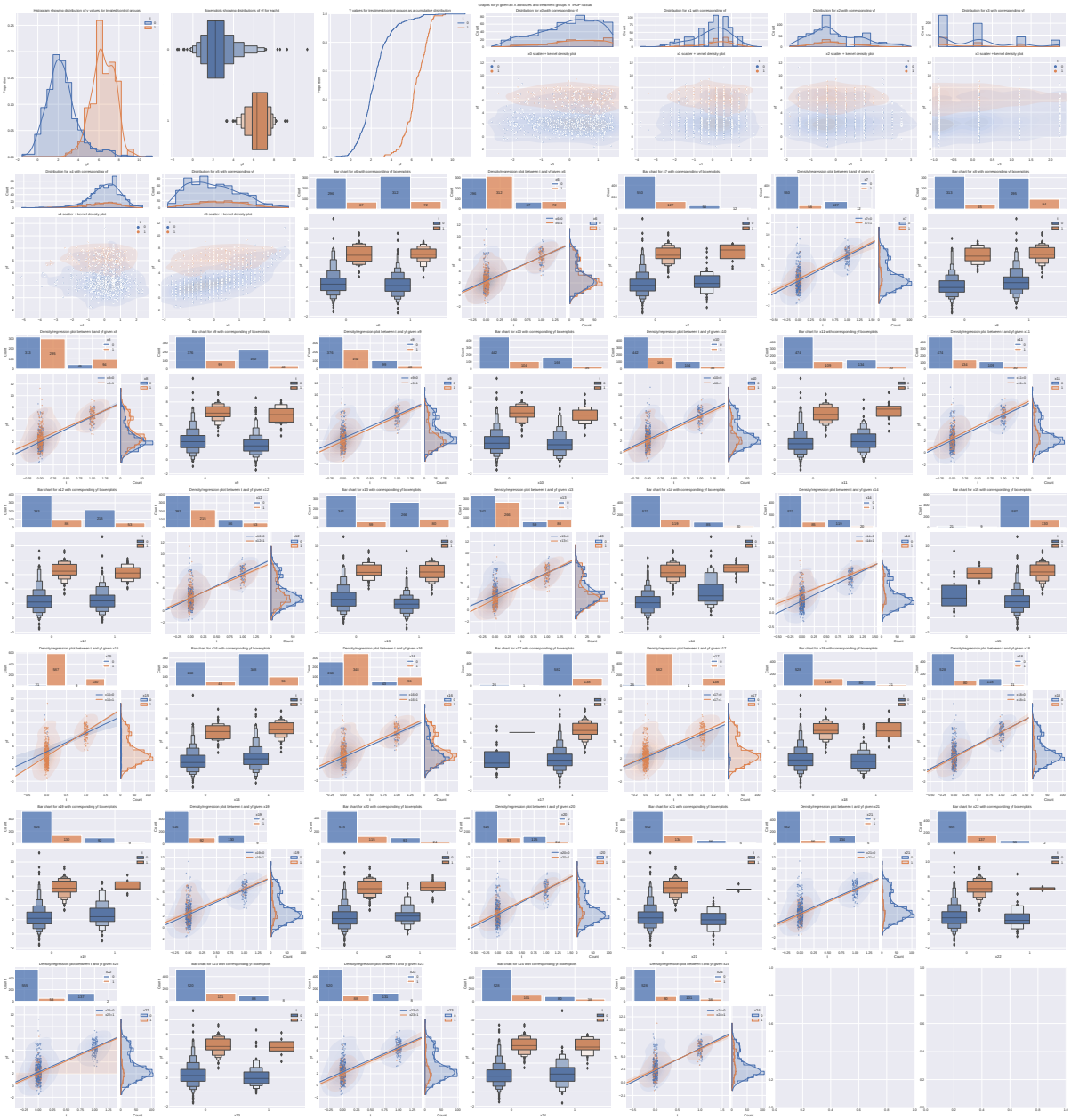

Figure 5: Graphs for the factual data in IHDP

These graphs are of the factual data (`yf`) within the IHDP dataset. These do illustrate the general positive correlation between individuals receiving treatment and having a higher `yf` value, but also shows how imbalanced this dataset is (most notably with `x17`, where only 27 individuals have `x17==0`.

These graphs clearly illustrate that the treated individuals generally have higher `yf` values than their untreated peers, with roughly similar interquartile ranges for treatment/control groups with different values for each binary `x` value (even in the extremely unbalanced cases).

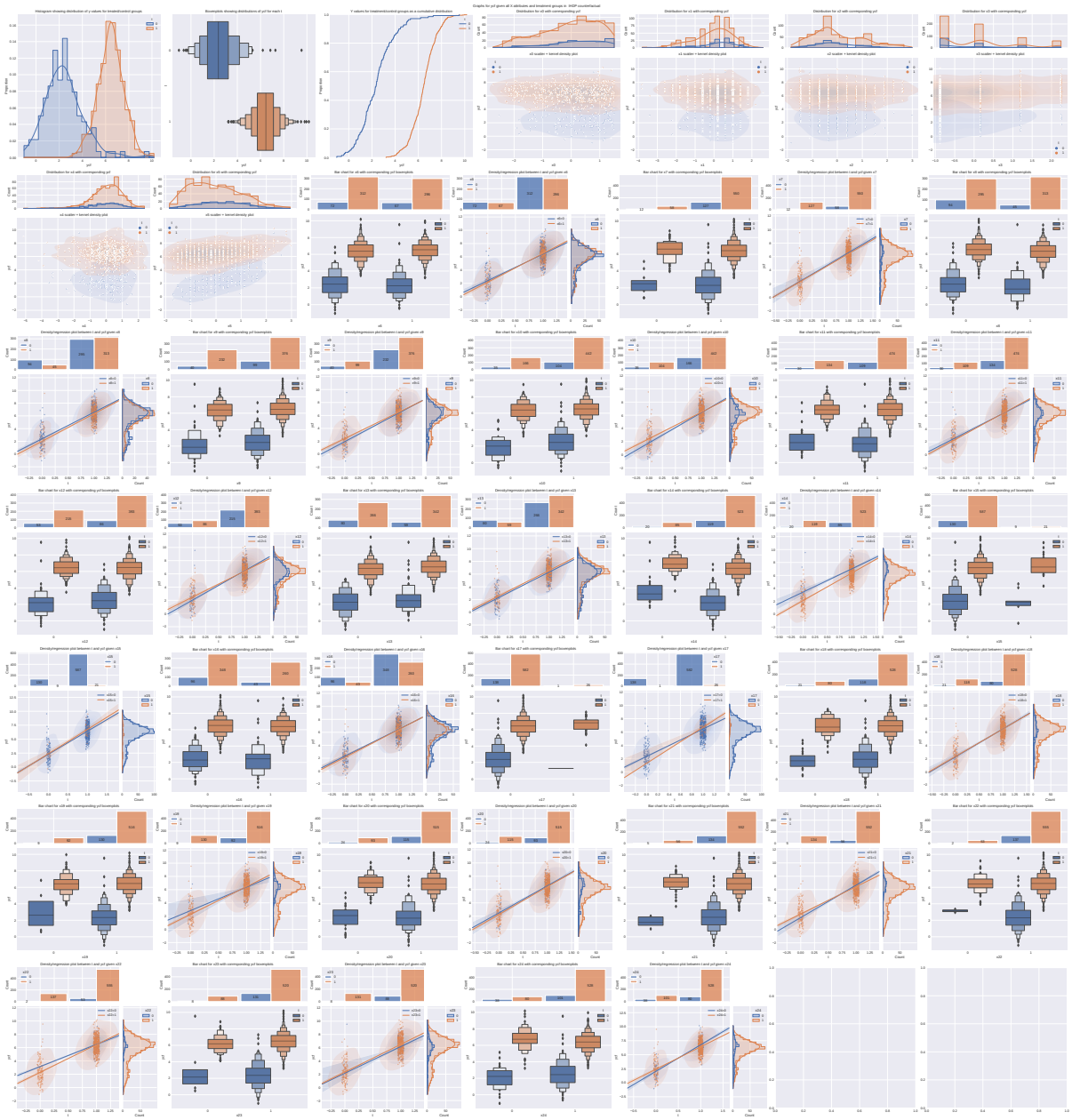

Figure 6: Graphs for the counterfactual data in IHDP

A.2 JOBS dataset visualizations

Here are the graphs for the JOBS dataset.

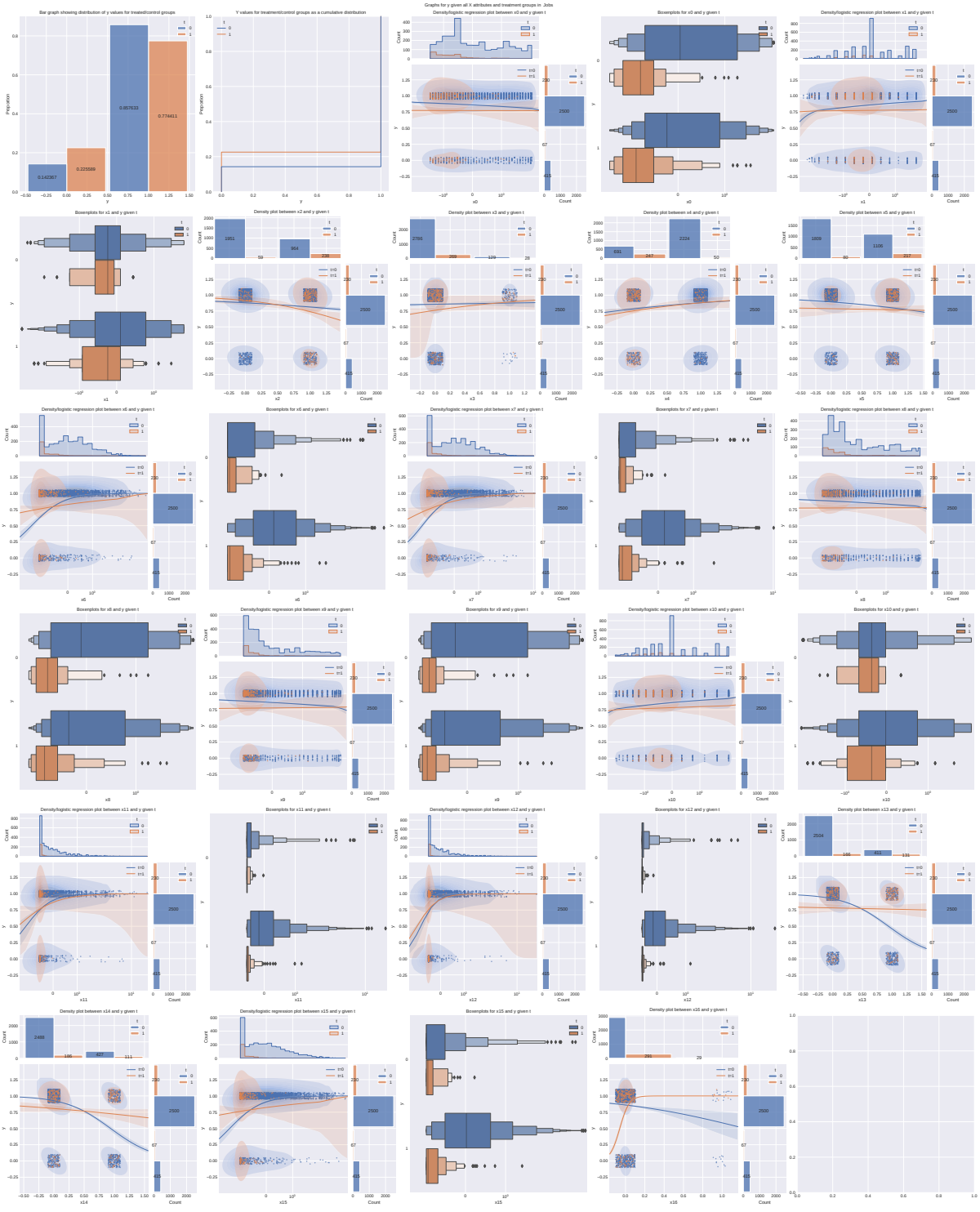

Figure 7: Several graphs for the JOBS dataset

References

[1] R. T. Gross *et al.*, *Infant health and development program (ihdp): Enhancing the outcomes of low birth weight, premature infants in the united states, 1985-1988*, 1993. DOI: 10.3886/ICPSR09795.v1. [Online]. Available: https://doi.org/10.3886/ICPSR09795.v1.

[2] J. Brooks-Gunn, F.-r. Liaw, and P. K. Klebanov, "Effects of early intervention on cognitive function of low birth weight preterm infants," *The Journal of Pediatrics*, vol. 120, no. 3, pp. 350–359, 1992, ISSN: 0022-3476. DOI: https://doi.org/10.1016/S0022-3476(05)80896-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022347605808960.

[3] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, vol. 76, no. 4, pp. 604–620, 1986, ISSN: 00028282. [Online]. Available: http://www.jstor.org/stable/1806062.

[4] R. H. Dehejia and S. Wahba, "Propensity score matching methods for non-experimental causal studies," National Bureau of Economic Research, Working Paper 6829, Dec. 1998. DOI: 10.3386/w6829. [Online]. Available: http://www.nber.org/papers/w6829.

[5] J. A. Smith and P. E. Todd, "Does matching overcome lalonde's critique of nonexperimental estimators?" *Journal of Econometrics*, vol. 125, no. 1, pp. 305–353, 2005, Experimental and non-experimental evaluation of economic policy and models, ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2004.04.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030440760400082X.

[6] J. Hill and E. A. Stuart, "Causal inference: Overview," in *International Encyclopedia of the Social Behavioral Sciences (Second Edition)*, J. D. Wright, Ed., Second Edition, Oxford: Elsevier, 2015, pp. 255–260, ISBN: 978-0-08-097087-5. DOI: https://doi.org/10.1016/B978-0-08-097086-8.42095-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080970868420957.

[7] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet, "Causal inference in public health," eng, *Annual review of public health*, vol. 34, pp. 61–75, 2013, PMC4079266[pmcid], ISSN: 1545-2093. DOI: 10.1146/annurev-publhealth-031811-124606. [Online]. Available: https://doi.org/10.1146/annurev-publhealth-031811-124606.

[8] C. Fernández-Loría and F. Provost, *Causal decision making and causal effect estimation are not the same... and why it matters*, 2021. arXiv: 2104.04103 [stat.ML].

[9] D. Machlanski, "Ce888: Data science and decision making lecture 4: Causal inference," University Lecture, Lecture delivered on: 2022–8-2, 2022.

[10] "sklearn.metrics.log$_{l}oss^{-}scikit - learn1.0.2documentation$." (Feb. 24, 2022), [Online]. Available: %5Curl%7Bhttps://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html#sklearn.metrics.log_loss%7D (visited on 02/24/2022).

[11] A. Matran-Fernandez, "Causal inference: Machine learning for causal inference from observational data," CE888 assignment brief, 2022.

[12] "sklearn.ensemble.RandomForestRegressor - scikit-learn 1.0.2 documentation." (Feb. 24, 2022), [Online]. Available: %5Curl%7Bhttps://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html?highlight=randomforestregressor#sklearn.ensemble.RandomForestRegressor%7D (visited on 02/24/2022).

[13] T. Fletcher, *Relevance vector machines explained*, 2010.

[14] "sklearn.linear$_{m}odel.ARDRegression^{-}scikit-learn1.0.2documentation$." (Feb. 23, 2022), [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ARDRegression.html#sklearn.linear_model.ARDRegression (visited on 02/24/2022).

[15] "sklearn.ensemble.AdaBoostRegressor - scikit-learn 1.0.2 documentation." (Feb. 23, 2022), [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html?highlight=adaboostregressor#sklearn.ensemble.AdaBoostRegressor (visited on 02/24/2022).

[16] "sklearn.metrics.r2$_{s}$core^{-}scikit − learn1.0.2documentation." (Feb. 23, 2022), [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (visited on 02/23/2022).

[17] "sklearn.linear$_{m}$odel.BayesianRidge^{-}scikit − learn1.0.2documentation." (Feb. 24, 2022), [Online]. Available: %5Curl%7Bhttps://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html#sklearn.linear_model.BayesianRidge%7D (visited on 02/24/2022).

[18] "econml.metalearners.XLearner - econml 0.13.0 documentation." (Feb. 14, 2022), [Online]. Available: https://econml.azurewebsites.net/_autosummary/econml.metalearners.XLearner.html (visited on 02/24/2022).

[19] "econml.dml.CausalForestDML - econml 0.13.0 documentation." (Feb. 14, 2022), [Online]. Available: %5Curl%7Bhttps://econml.azurewebsites.net/_autosummary/econml.dml.CausalForestDML.html%7D (visited on 02/24/2022).

[20] "econml.dr.ForestDRLearner - econml 0.13.0 documentation." (Feb. 14, 2022), [Online]. Available: %5Curl%7Bhttps://econml.azurewebsites.net/_autosummary/econml.dr.ForestDRLearner.html%7D (visited on 02/24/2022).

[21] K. Battocchi, E. Dillon, M. Hei, *et al.*, *Econml: A python package for ml-based heterogeneous treatment effects estimation*, https://github.com/microsoft/EconML, Version 0.x, 2019.