# Case Study IV, Interim Report II

*Jake Epstein, Daniel Spottiswood, Man-Lin Hsiao, Michael Tan, Sahil Patel*

*11/01/2019*

## Introduction

The goal of this case study is to evaluate the causal impact of debit card ownership on household spending. The data come from the Italy Survey on Household Income and Wealth (SHIW), a 1995-1998 survey of 584 Italian households. The dataset includes 1995 and 1998 monthly household spending, whether the household had exactly one debit card in 1998 and demographic information including family size, geographic region and average age. In this report, we will create a model to estimate the causal impact of debit card ownership on household spending, utilizing propensity score methods to ensure model balance.

## Exploratory Data Analysis

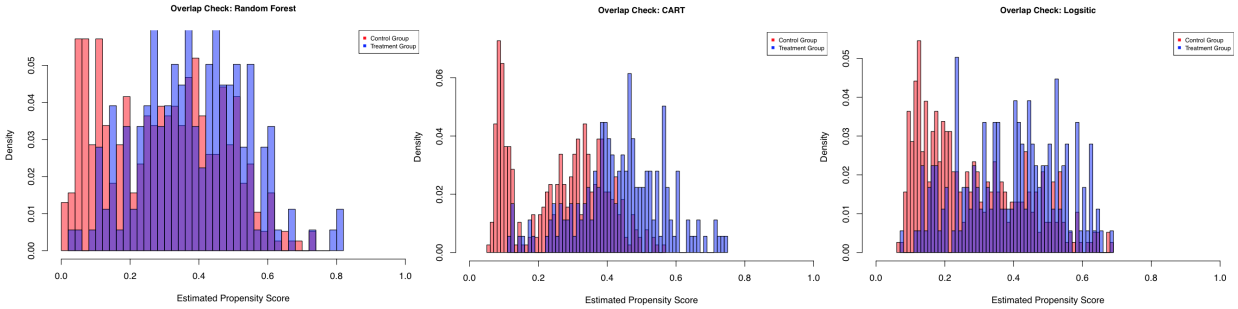
We begin our exploratory data analysis by looking at spending. In 1995 and 1998, households with debit cards tended to spend more than households without. The distribution of difference in household spending is centered at around 0, indicating most households spent about the same amount in 1998 as they did in 1995. The distribution of changes for households with debit cards has slighly more weight on the positive side, indicating that these households may have increased their spending slightly relative to non-debit card households. We also looked into spending as a percentage of income or of wealth, and the results were consistent with those above.

We then examined relationships between demographic characteristics and 1998 spending. Some selected plots are shown above. Our initial analysis indicates that family size is positively associated with spending, which is intuitive given the cost of raising children. Additionally, families with household heads who have higher educational status tend to spend more than those headed by less educated individuals. This may be a function of income or wealth, as higher educated individuals tend to earn more; regardless, it is worth exploring further. Finally, both income and wealth are positively associated with spending, and households with debit cards tend to spend more at all levels of income and wealth.

We move forward with attempts to balance the covariates using matching, weighting and propensity scores as these plots illustrate significant differences in the covariates between the treatment and control groups. y = "1998 Spending (% of Income)", x = "Geography")

## Data Balancing

To examine data balance between control and treatment groups, we looked at three different propensity score models: logistic regression, random forest, and CART, a decision-tree based machine learning model. The overlap in propensity scores for each model is shown below. Logistic regression and random forest show good overlap, whereas CART does not. Given that logistic and random forest perform similarly, and logistic regression is a simpler and more interpretable model, we will use the logistic regression model for propensity scores.

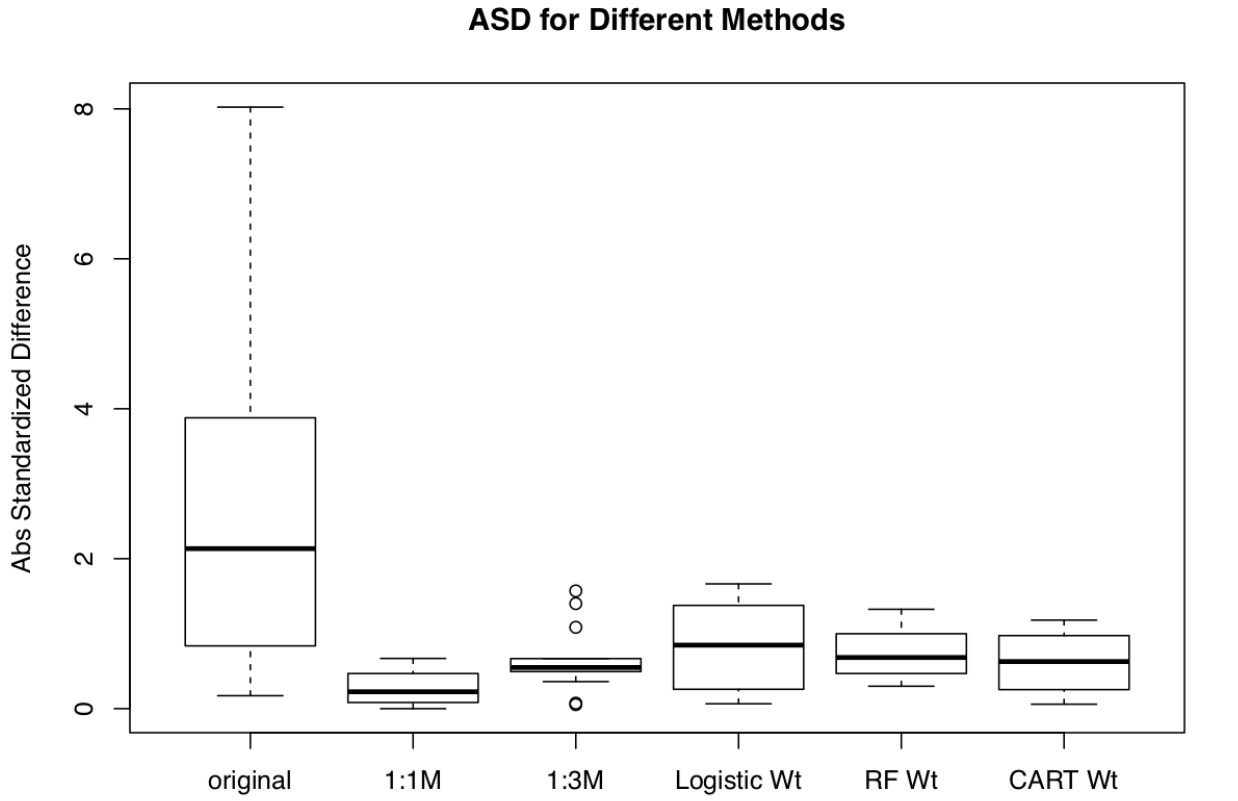



[TODO: Get rid of excess outputhistograms, etc]

[TODO: Fix all loveplots, focus on matching]

[TODO: Add lines, categorical variables to matching love plots]

[TODO: Discussion of why we're choosing 1:1 matching for regression adjustment]




Moving forward with 1:1 matching, we see that the absolute standardised differences are improved significantly across almost all covariates.

## Regression Adjustment

We attempt to fit models on top of the 1:1 matching with replacement. We first use a basic linear model before fitting more flexible models, random forest and extreme gradient boosting.

The motivation for regression adjustment is to impute missing potential outcomes by fitting models to the unobserved counterparts of observed data subsets. We have elected to include a regression adjustment in addition to our propensity score matching as it allows us to mitigate the sensitivity of our models to the model specification used to impute these unobsereved values, specifically as it pertains to the covariate imbalance. As described above, our 1:1 matched data has demonstrated the most compelling distribution of absolute standardised difference, and therefore yields the most strongest results for the overlap of data through a propensity score methodology. Thus we partitioned the data into treatment and control groups through the 1:1 mathcing method, and then imputed unobserved control values for the observed treatment values, and vice versa for the observed control values. We began testing our models by running a simple linear (OLS) model on the original unmatched dataset. The results of this test proved unfruitful as the ATT for this model was lower than the corresponding ATE. This suggests that the effect of individuals who were observed to own debit cards in 1998 spent on average 185 more than those who did not. Furthermore, the unmatched model suggests that after imputing control values for those who did own a debit card in 1998, that the implied treatment effect increased to 218 in spending. Ergo, the model implies that those who recieved the treatment would have spent less than those in the corresponding observed control group. Intutively, this reinforces our intuition to use the matched data as it would allow our model to account for unoberserved confounders better, and therfore allowing for stronger causal inference As illustrated by the table above, the ATT is higher than the ATE across all models that were built with the 1:1 matched data. Intutively, this suggests that matching enabled us to obviate a case of Simpson's Paradox as the unmatched model demonstrated the opposite intuition. In terms of the flexibility of our models, the machine learning models (the gradient boost and random forest) are intrinsically more flexible than our OLS model as they learn from previous errors by design. That being said, it somewhat cumbersome to ascertain which of our machine learning models is in fact more flexible. The forest is useful as it is easier to fit as it only depends on the number of trees and the number of factors selected; in addition, the forest does not overfit the data as much as the boosted model. Nevertheless, the boosted model is also compelling as it is capable of optimising a broader selection of objective function than the forest, and typically leads to more accuracy with fewer trees. The complexity of the data at hand has led us to prefer our boosted model, which yielded the highest ATT (286) and ATE (169).

In order to estimate the error and significance of our models for ATT and ATE, we generate 500 bootstrap samples of the data,. Using the bootstrap samples, we recalculated the propensity scores and estimated the resulting causal effects for each dataset accordingly. In our bootstrapping process, we evaluated both the ATT and ATE for weighting, weighting with double robustness, 1:1 matching with our linear model, and 1:1 matching with our gradient boosted model.

## Conclusions

| | R Adj. w/out Match | Weighting | Double Robustness | R Adj. Linear | R Adj. XGBoost |
|---|---|---|---|---|---|
| ATE | 218.937029 | 112.097080 | 126.724605 | 127.194161 | 169.611631 |
| ATE SD | 33.664420 | 50.409745 | 65.170615 | 78.589876 | 50.909140 |
| ATT | 185.353691 | 67.257699 | 258.143468 | 218.624545 | 286.750533 |
| ATT SD | 64.188673 | 58.220035 | 85.325756 | 58.132368 | 71.830614 |
| ATT/AOT | 0.091 | 0.033 | 0.13 | 0.11 | 0.14 |
| P-Value | 0.0019 | 0.12 | 0.0012 | 0.000085 | 0.000033 |

The results of bootstrapping support our aforementioned belief in the gradient boosted model. As portrayed in the table above, the boosted model was the most significant (as estimated by the model's p-value) and has the lowest estimated (via bootstrapping) standard deviation for ATE across all models explored. Interestingly, the boosted model did have the highest bootstrap estimate for ATE standard deviation, however given the

desirability of the model's other metrics and the flexibility of the model holistically, we continue to endorse the gradient boosted model as our model of choice. Overall, our investigation into the effects of debit card ownership on spending has demonstrated that there is in fact a positive treatment effect associated with owning a debit card in 1998. While it may prove difficult to extrapolate the results of this study to other geographies and time periods, we believe that through implementing 1:1 matching and regression adjustments with flexible machine learning models we have achieved the ability to draw some causal inference. Namely, we estimate the ATT to be approximately 286 and the ATE to be approximately 169.