Case Study IV, Interim Report II

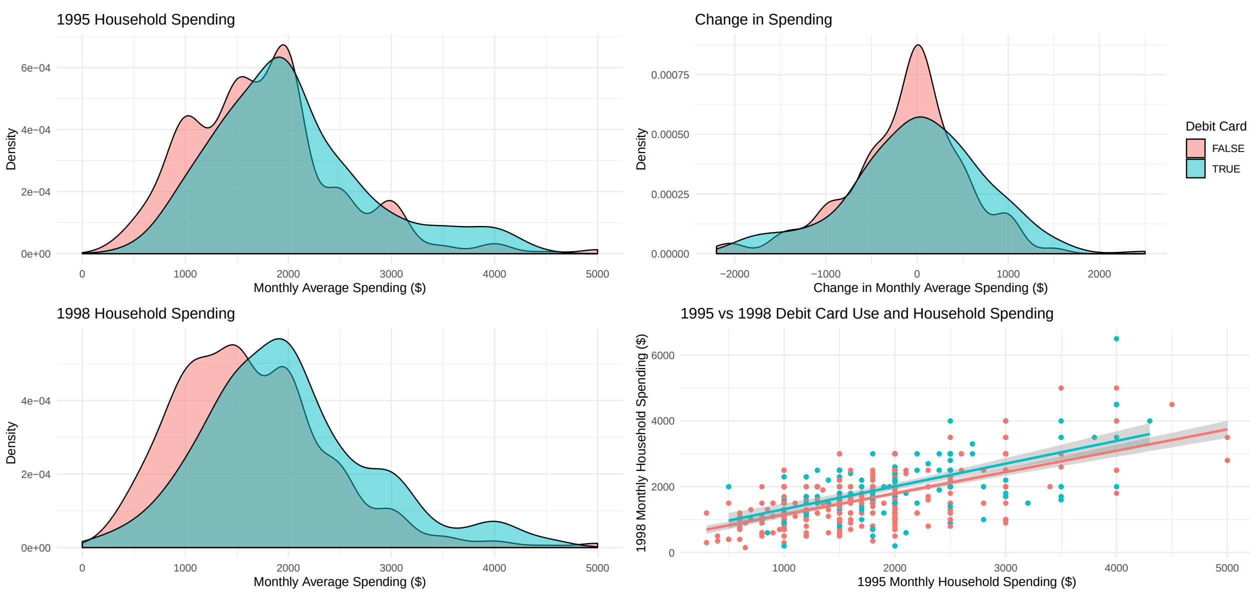
Jake Epstein, Daniel Spottiswood, Man-Lin Hsiao, Michael Tan, Sahil Patel

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



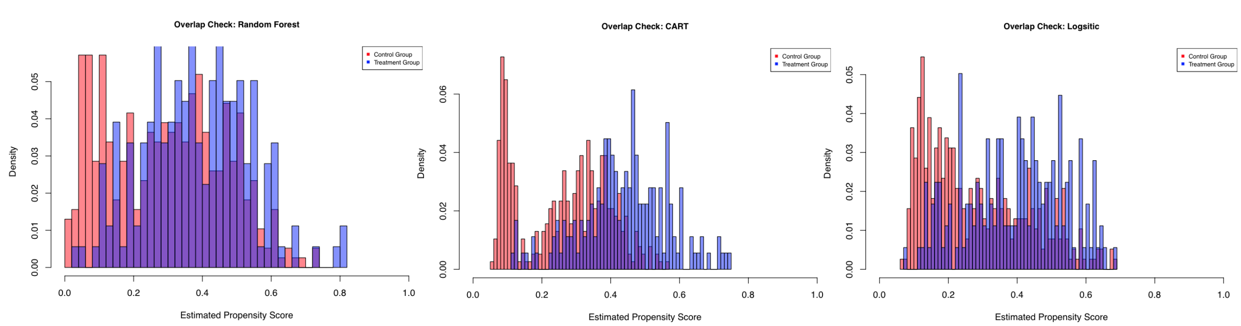
We begin our exploratory data analysis by looking at spending. In 1995 and 1998, households with debit cards tended to spend more than households without. The distribution of difference in household spending is centered at around 0, indicating most households spent about the same amount in 1998 as they 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.

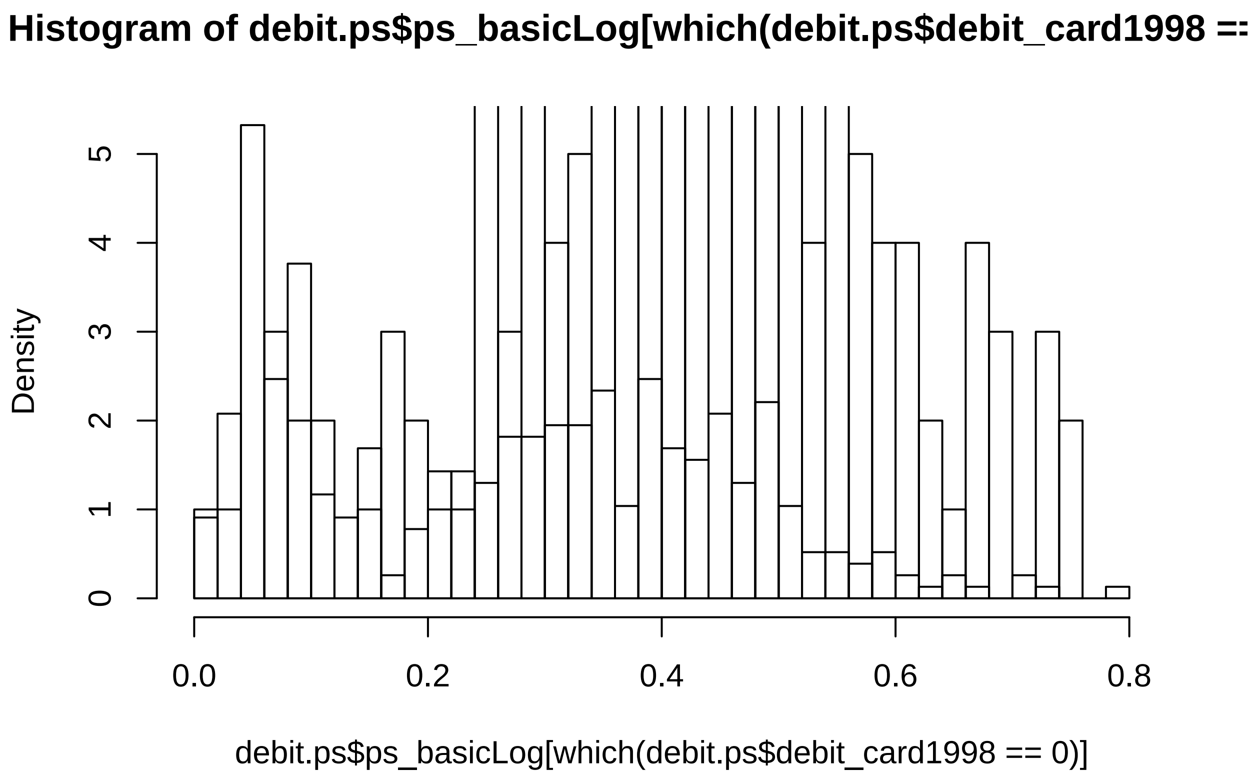
We move forward with attempts to balance the covariates using matching, weighting and propensity scores as these plots illustrate significant differences in the covariates between the treatment and control groups. y = "1998 Spending (% of Income)", x = "Geography")

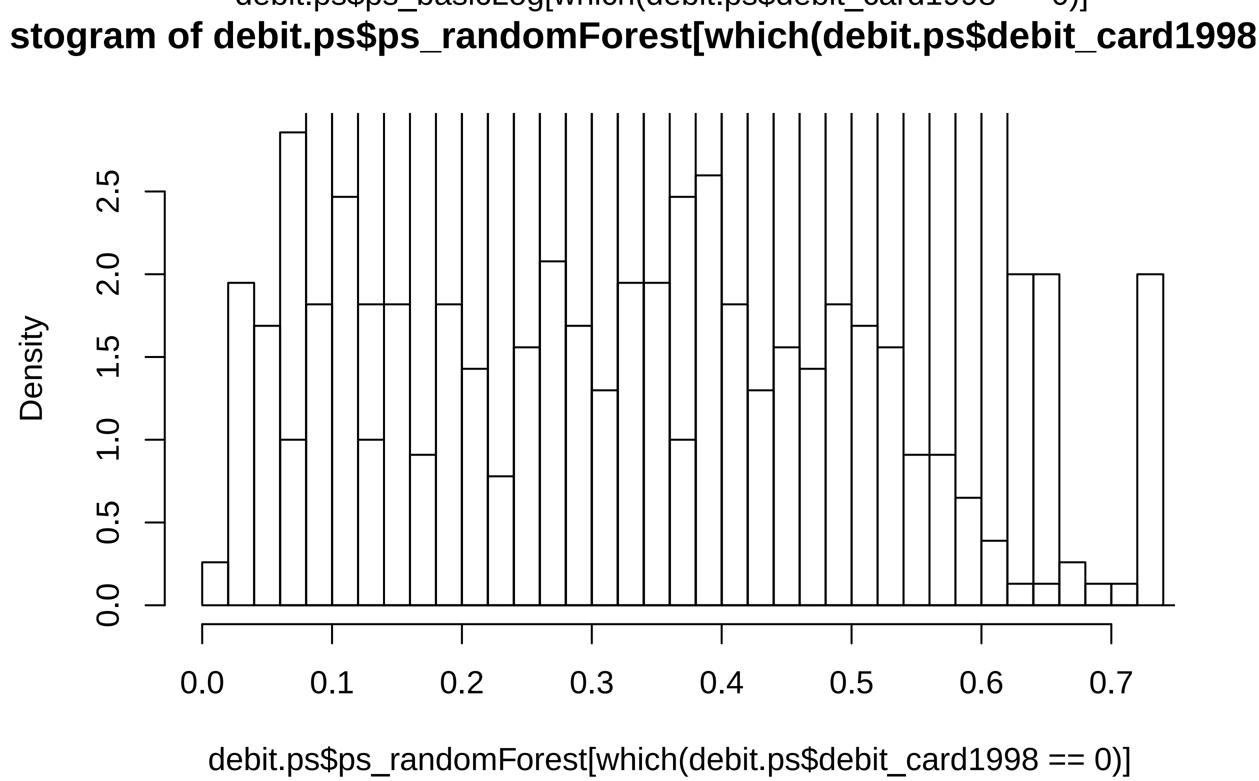
Data Balancing

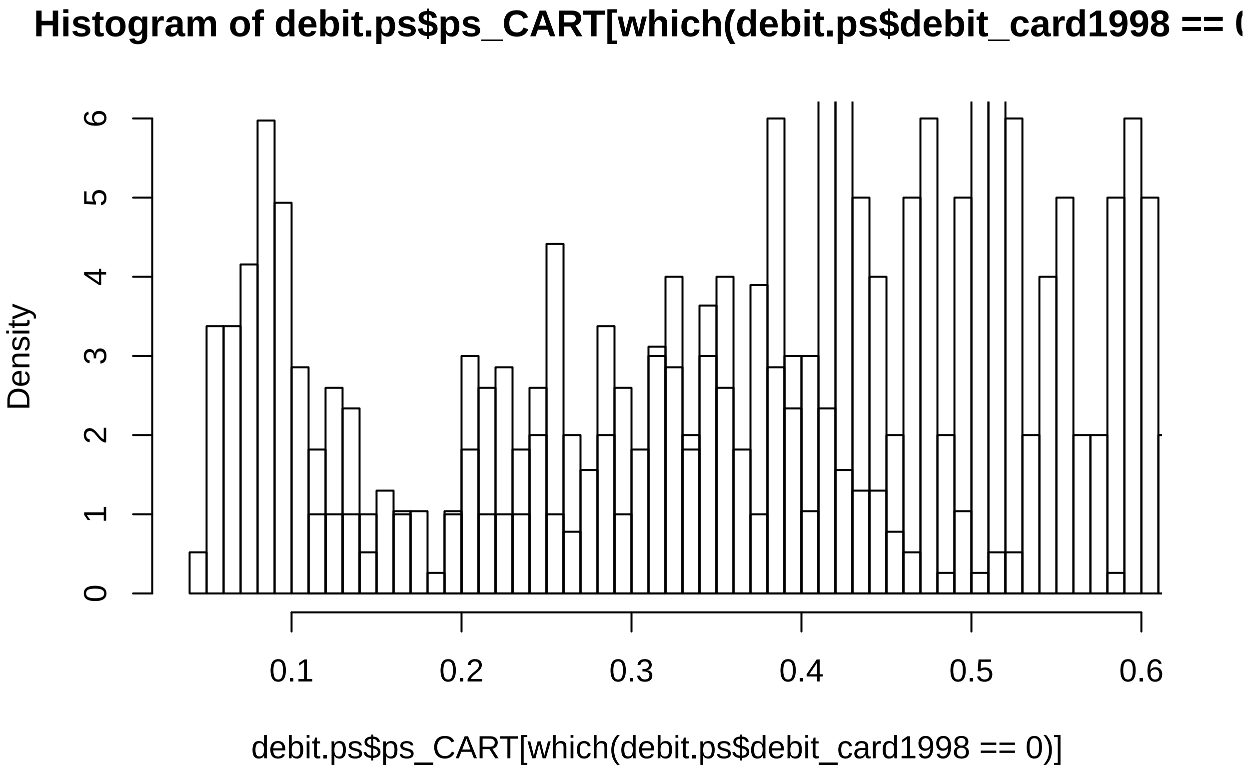
To examine data balance between control and treatment groups, we looked at three different propensity score models: logistic regression, random forest, and CART, a decision-tree based machine learning model. The overlap in propensity scores for each model is shown below. Logistic regression and random 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.



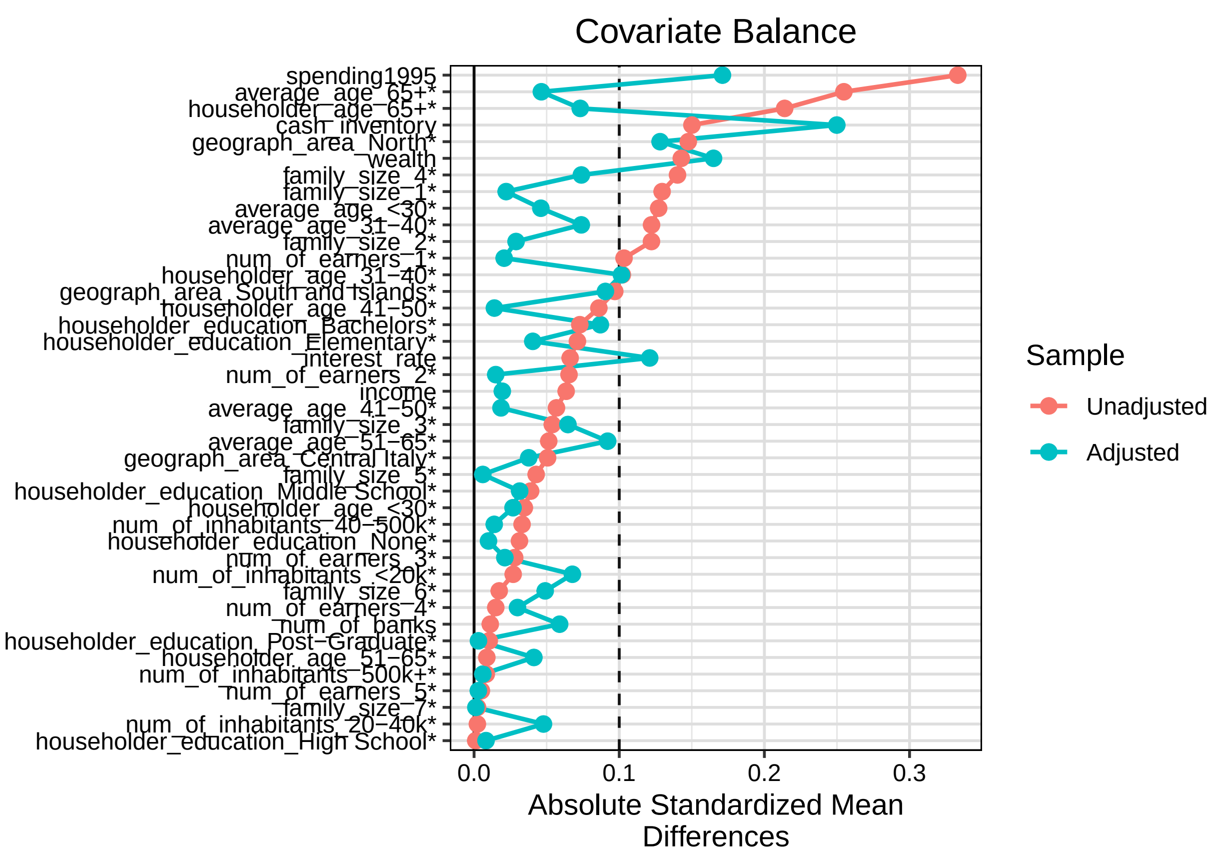
[TODO: Get rid of excess outputhistograms, etc]

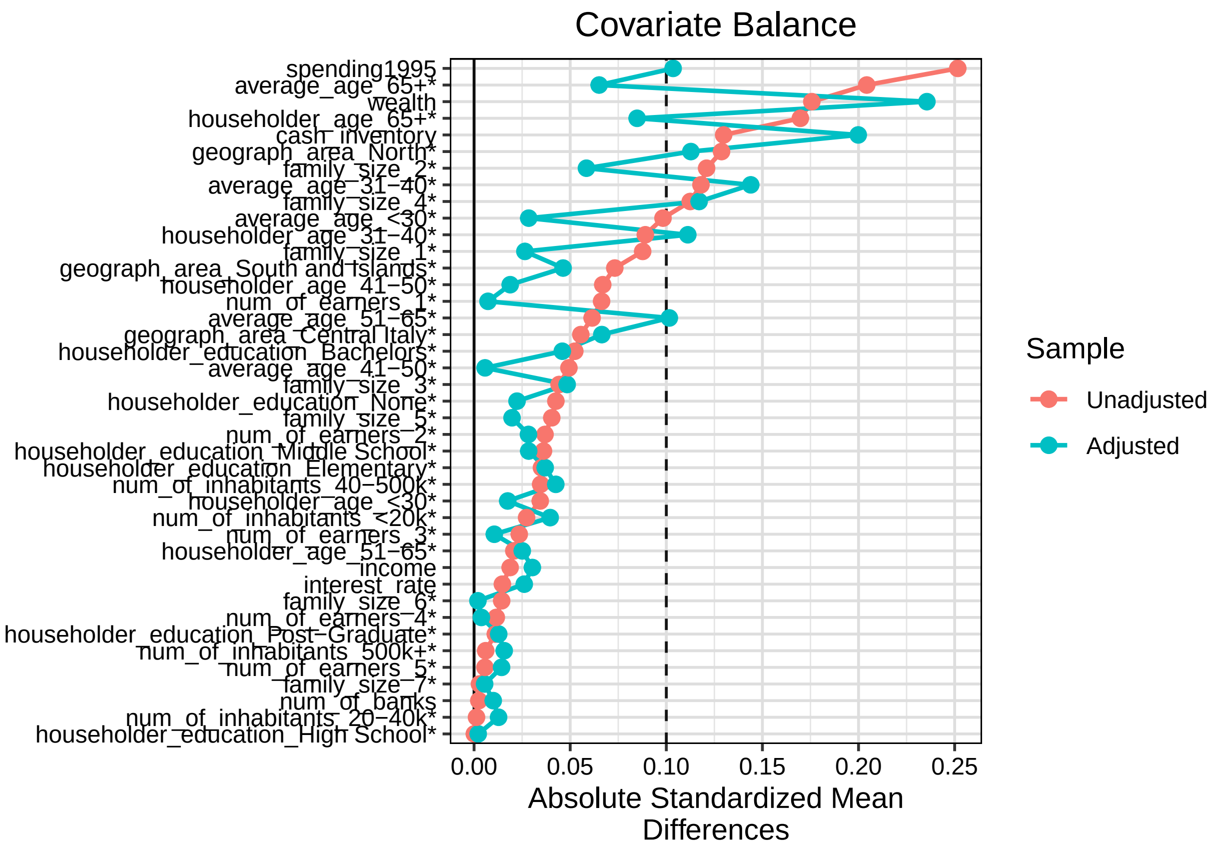

Histogram of debit.ps$ps_basicLog[which(debit.ps$debit_card1998 ==
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
debit.ps$ps_basicLog[which(debit.ps$debit_card1998 == 0)]


stogram of debit.ps$ps_randomForest[which(debit.ps$debit_card1998
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
debit.ps$ps_randomForest[which(debit.ps$debit_card1998 == 0)]

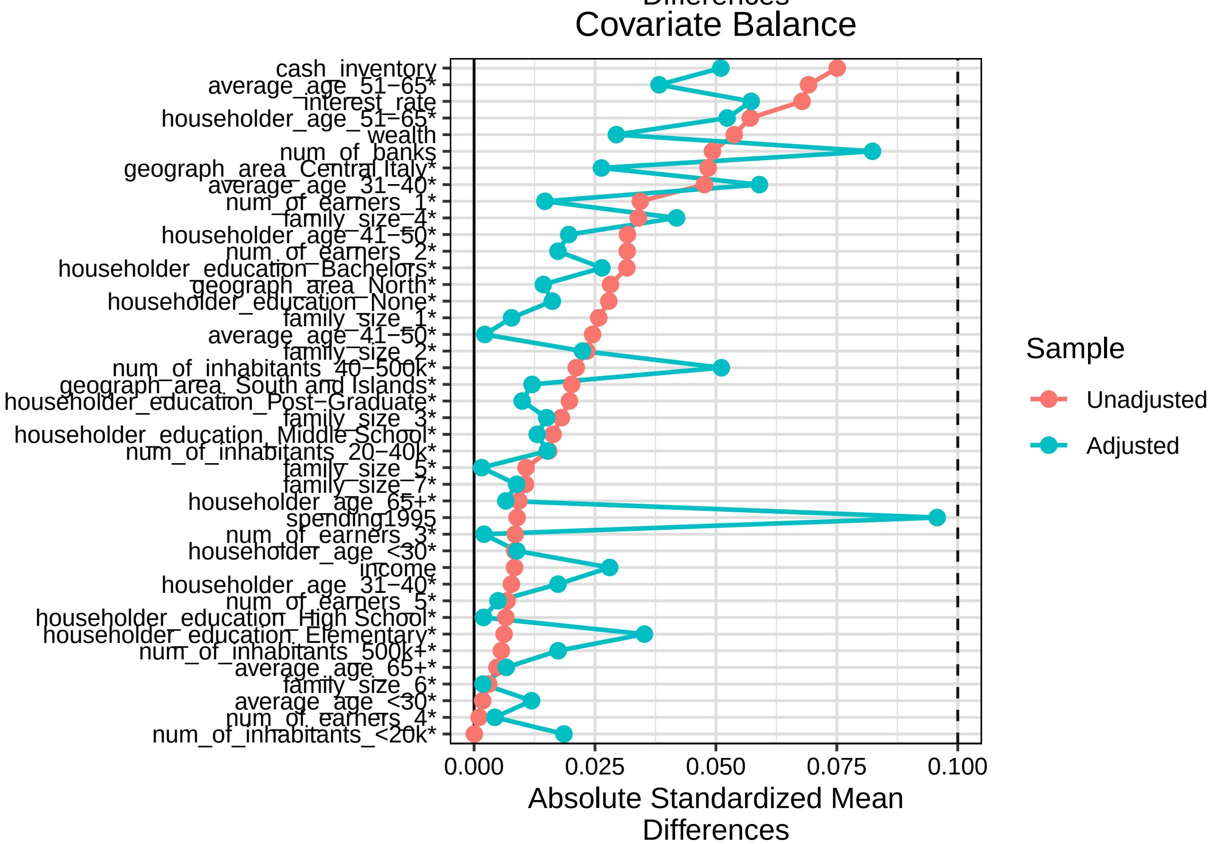


[TODO: Fix all loveplots, focus on matching]

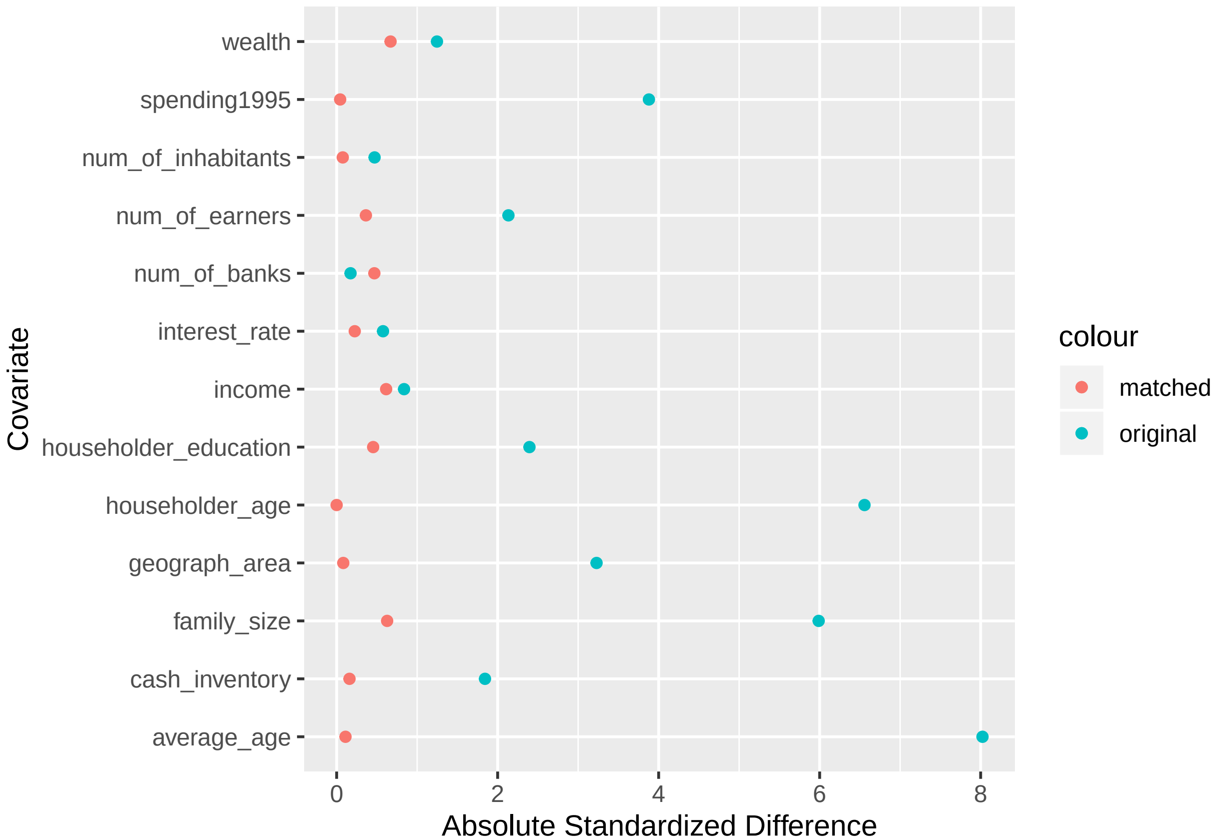




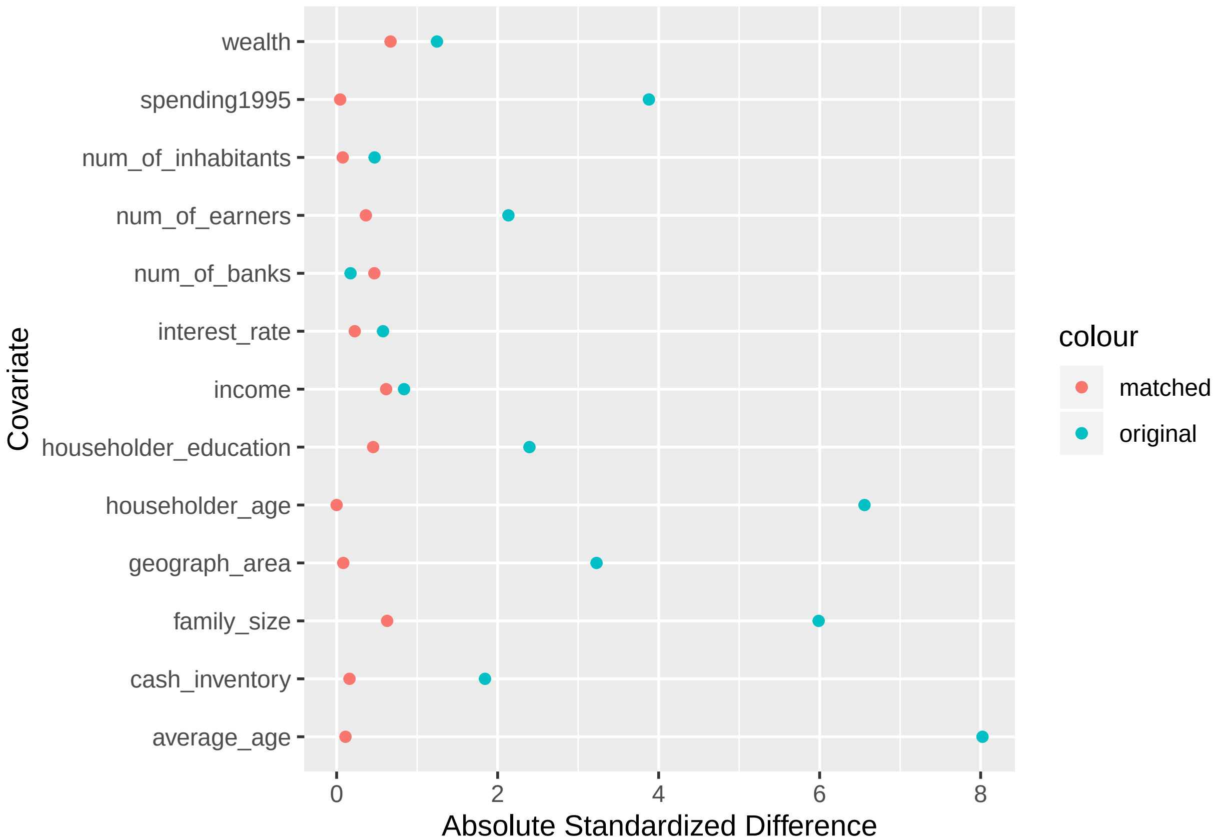
Covariate Balance
spending1995
average_age_65+*
wealth
householder_age_65+*
cash_inventory
geograph_area_North*
family_size_2*
average_age_31−40*
family_size_4*
average_age_<30*
householder_age_31−40*
family_size_1*
geograph_area_South and Islands*
householder_age_41−50*
num_of_earners_1*
average_age_51−65*
geograph_area_Central Italy*
householder_education_Bachelors*
average_age_41−50*
family_size_3*
householder_education_None*
family_size_5*
num_of_earners_2*
householder_education_Middle School*
householder_education_Elementary*
num_of_inhabitants_40−500k*
householder_age_<30*
num_of_inhabitants_<20k*
num_of_earners_3*
householder_age_51−65*
income
interest_rate
family_size_6*
num_of_earners_4*
householder_education_Post−Graduate*
num_of_inhabitants_500k+*
num_of_earners_5*
family_size_7*
num_of_banks
num_of_inhabitants_20−40k*
householder_education_High School*
0.00 0.05 0.10 0.15 0.20 0.25
Absolute Standardized Mean Differences
Sample
Unadjusted
Adjusted



Covariate Balance
cash_inventory
average_age_51−65*
interest_rate
householder_age_51−65*
wealth
num_of_banks
geograph_area_Central Italy*
average_age_31−40*
num_of_earners_1*
family_size_4*
householder_age_41−50*
num_of_earners_2*
householder_education_Bachelors*
geograph_area_North*
householder_education_None*
family_size_1*
average_age_41−50*
family_size_2*
num_of_inhabitants_40−500k*
geograph_area_South and Islands*
householder_education_Post−Graduate*
family_size_3*
householder_education_Middle School*
num_of_inhabitants_20−40k*
family_size_5*
family_size_7*
householder_age_65+*
spending1995
num_of_earners_3*
householder_age_<30*
income
householder_age_31−40*
num_of_earners_5*
householder_education_High School*
householder_education_Elementary*
num_of_inhabitants_500k+*
average_age_65+*
family_size_6*
average_age_<30*
num_of_earners_4*
num_of_inhabitants_<20k*
0.000 0.025 0.050 0.075 0.100
Absolute Standardized Mean Differences
Sample
Unadjusted
Adjusted

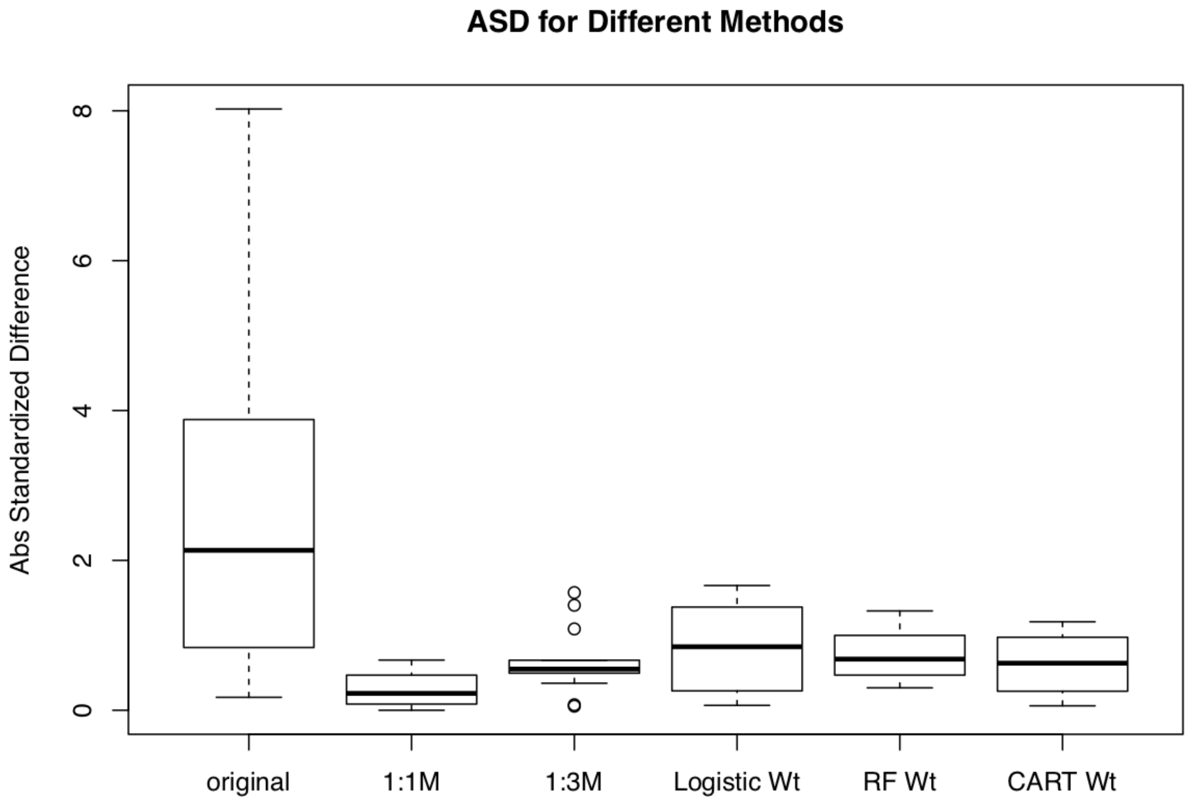

wealth
spending1995
num_of_inhabitants
num_of_earners
num_of_banks
interest_rate
income
householder_education
householder_age
geograph_area
family_size
cash_inventory
average_age
Covariate
0
2
4
6
8
Absolute Standardized Difference
colour
matched
original

```
## pdf
##   2
```



[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 mean 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.

```
##    ATT_mix  ATE_mix ATT_mix_boost ATE_mix_boost ATT_mix_forest
## 1 218.6245 127.1942      286.7505      169.6116       254.7947
##   ATE_mix_forest
## 1       139.7986
```

[TODO: Writeup bootstrapping]

```
## % latex table generated in R 3.5.3 by xtable 1.8-4 package
## % Mon Nov  4 18:46:59 2019
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & mean & V2 & V3 & V4 \\
##   \hline
## 1 & 112.0971 & 126.7246 & 127.1942 & 169.6116 \\
##   2 & 51.4313 & 68.4296 & 104.0844 & 55.0839 \\
```

```
##   3 & 169.0854 & 258.1435 & 218.6245 & 286.7505 \\
##   4 & 66.3133 & 71.7979 & 71.6663 & 78.6000 \\
##    \hline
## \end{tabular}
## \end{table}
```

Conclusions

	Weighting	Double Robustness	Reg Adj. Linear	Reg Adj. XGBoost
ATE	112.0971	126.7246	127.1942	169.6116
ATE SD	53.9495	65.9373	86.2670	52.9248
ATT	67.2577	258.1435	218.6245	286.7505
ATT SD	61.9402	70.5002	67.2079	73.0510
ATT/AOT	0.03317021	0.12731142	0.10782145	0.14141988
P-Value	.14	.00013	.000571	.0000433

[TODO: Add interpretations and conclusions of above table]