UNIVERSITY OF MANNHEIM

MASTER'S THESIS

The Impact of Transitory Income Shocks on Labor Supply Decisions: A Machine Learning Approach

---

Author:
Dominik Lauer

Student ID:
1408215

Supervisor:
Prof. Krzysztof Pytka, Ph.D.

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

at the

Chair of Quantitative Macroeconomics
Department of Economics

July 27, 2020

Declaration of Academic Integrity

I, Dominik Lauer, hereby declare that the master's thesis entitled "The Impact of Transitory Income Shocks on Labor Supply Decisions: A Machine Learning Approach" is entirely my own work, and that I have employed no sources or aids other than the ones listed. I have clearly marked and acknowledged all ideas and illustrations that have been taken directly or indirectly from the works of others. I also confirm that this thesis has not been submitted in this form, or a similar form, to any other academic institution.

Signed:

_____

Date:

_____

UNIVERSITY OF MANNHEIM

Department of Economics

The Impact of Transitory Income Shocks on Labor Supply Decisions: A Machine Learning Approach

by Dominik Lauer

Abstract

I study the effect of transitory income shocks on individual household labor supply decisions at the intensive margin. While there have already been several studies focusing on households' consumption choices upon receiving transitory income in the form of tax rebates, empirical research on the impact on labor supply decisions has been widely ignored. To address this issue, I use a machine learning approach. While finding no long-term impact of rebate payments on labor income, my results suggest that each rebate dollar reduces households' labor income by around 9 cents in the month of rebate payment. This corresponds to an average working hour reduction of around 4 hours in rebate months. I also find smaller lagged responses of income reduction. Households' reactions are widely heterogeneous, with young and single households as well as households paid by the hour being main drivers. However, due to their high variance, my estimates generally show only a limited statistical significance.

Contents

List of Figures

List of Tables

1 Introduction

Fiscal stimulus packages have been used repeatedly in the past to steer countries through economic crises. In addition to monetary measures, fiscal packages play a decisive role in the demand-side stimulation of economic growth. Whether fiscal stimulus packages actually achieve their desired stimulus effect in reality, or whether governmental financial aids do not lead to a significant mobilization of economic performance, depends largely on the targeting of these measures. In recent years, a large number of researchers have been investigating the extent to which fiscal stimulus packages have motivated individuals and households to increase consumption - and consequently economic growth. Starting with the measurement of average changes in household consumer behavior, the focus of the analysis in the recent past has been primarily on the heterogeneity of consumption reactions. With the foundations being laid in the 1950s, where Modigliani and Brumberg (1955) and Friedman (1957) published the life cycle hypothesis (LCH) and the permanent income hypothesis (PIH) respectively, recent literature mainly focuses on heterogeneous marginal propensities to consume (MPCs), which measure the proportion of an increase in disposable income an individual spends on consumption (amongst others, Misra and Surico (2014)). Kaplan and Violante (2014) study the heterogeneity of MPCs using a structural model framework with two assets, thereby generating two groups of households with high MPCs. More recently, Carroll et al. (2017) find that heterogeneous preferences in the form of different levels of impatience results in significantly heterogeneous MPCs.

However, one important element in the analysis of fiscal stimulus programs has been largely ignored in these studies. To the best of my knowledge, Powell (2020) is the first scholar who devotes much attention to the heterogeneous impact of tax rebates on household labor supply decisions. Powell studies the 2008 economic stimulus payments (ESPs) and finds that households temporarily reduce labor income when receiving a tax rebate. As he goes on to explain, this reduction is largely voluntary, which means that the consumption stimulus provided by fiscal packages is partly offset by the decline in labor input. By estimating not only statistically significant but also high absolute reductions in production capacity, Powell underscores the importance of this effect for the first

time.

At this point my Master's thesis comes in. The thesis combines Powell's question about the effects of tax rebates on households' work decisions with a proven and effective method for determining heterogeneous treatment effects. Thus, this thesis tries to not only answer the question of whether and how much households changed their labor income - and hence their individual labor supply - in the aftermath of the receipt of tax rebates due to the economic stimulus payments in 2008, but also to identify the key drivers of these changes as well as heterogeneous responses for different household characteristics. In contrast to Powell, I will focus on households at the intensive margin only, excluding unemployed households with a reported labor income of zero dollars from the analysis. Unlike many other analyses that tried to estimate treatment effects in the data, this thesis does not use a regression estimation but a random forest machine learning algorithm to predict the change in labor income due to the receipt of tax rebates. To the best of my knowledge, this thesis is the first one that tries to measure labor income changes due to transitory income shocks using a machine learning approach. In brief, it allows to document heterogeneity in labor income changes without relying on assumptions as in structural models (Gulyas and Pytka, 2019). Advantages and disadvantages of this approach will be examined in more detail in the further course of the thesis.

The thesis is organized as follows. The next section gives an overview about literature that focuses on household labor supply theory and already existing papers on changes in the supply of labor. This is followed by the provision of some background information on the 2008 Economic Stimulus Act and its rebate payments. Chapter 4 includes the description of the data as well as the empirical approach. The major part of this chapter is devoted to the theory of the Random Forest algorithm and the description of the uplifting approach used for the estimations. While the main findings and the heterogeneity analysis are presented in chapter 5, I use chapter 6 for a conclusion.

2 Literature Review

This section aims to give the reader a brief overview of labor supply theory at the beginning.[1] More recent findings on a variety of influences on individuals' and households' labor supply decisions are discussed in more detail. Also, the results of Powell (2020) will be described, as this work is largely inspired by Powell's paper.

History of Labor Supply Theory Analyses of labor supply theory date back to the 1960s, when several scholars began to explore individuals' motivations to take up work and the effects of public policies on labor determinants. From the very beginning, economists attributed great importance to the precise study of labor market economic effects. In fact, the impact of governmental programs on individuals' employment and hours of work is often crucial when it comes to the design of public policies (Blundell and MaCurdy, 1999). To my knowledge, Becker (1965) is one of the first to incorporate non-working "leisure" time into the individual's utility maximization problem by adding a separate time constraint to the already existing budget constraint. Due to the implementation of leisure and work time into the utility maximization, a microfoundation for changes in the aggregate labor supply was laid. Already before, Mincer (1960) examines different components of family income and concludes that these reflect family labor choices and vice versa. In the late 1960s, economists started to distinguish between labor supply effects on the overall labor participation rate of households (also referred to as the *extensive margin*) and on the hours worked (*intensive margin*). This distinction proved to be groundbreaking in order to explore the consequences and reasons for the so-called self-selection bias in the following years. Amongst others, DaVanzo et al. (1973), Borjas and Heckman (1978) and Juhn et al. (1991) underline that participation decisions are much more affected by changes in wages or income than adjustment effects in working hours for those already in employment. A large number of other areas of labor supply theory has since been investigated. For instance, the effects of taxation on labor supply have been studied intensively (Kosters, 1967; MaCurdy et al., 1990), entry and exit decisions of households as reactions to changes in wages or income have been examined

[1] Blundell and MaCurdy (2008) is recommended for a general overview.

more closely (Coleman, 1987; Alogoskoufis, 1987; Blundell et al., 2011), and the static theory of individual labor supply has been extended to several periods, leading to the life-cycle labor supply theory and the intertemporal substitution hypothesis (Heckman and Macurdy, 1980; MaCurdy, 1981; Owen, 1989; Card, 1991).[2] In the recent past, and due to the usage of better computational methods, the focus of labor supply research has shifted towards more empirical and data-driven analyses. In particular, the development of refined analyses and statistical procedures has made large amounts of data more accessible for evaluations in greater detail.

Labor Supply Responses to Transitory Shocks In recent years, there has been an accumulation of studies on the effects of temporary changes in various economic variables on labor supply. Camerer et al. (1997) studies the effect of transitory changes in wages on the labor supply of cabdrivers in New York City by estimating wage elasticities. His findings suggest that cabdrivers seem to make labor supply decisions on a daily basis rather than intertemporally substituting labor and leisure across days. Using a slightly different statistical approach, Farber (2005) contradicts Camerer's results by showing that daily income effects for cabdrivers are small and that their labor supply behavior is consistent with the standard neoclassical model. By examining the response of Florida lobster fishermen to temporary changes in income, Stafford (2015) argues that Camerer's criticism of the neoclassical model of labor supply is based on statistical measurement errors rather than on actual false model considerations. She finds positive and significant wage elasticities of daily working hours and participation rate, indicating that fishermen increase their work time or even start working when earnings are temporarily high. A somewhat other analysis is carried out by Fehr and Goette (2007). They use an "ideal" data set to study workers' responses to transitory wage changes, which they construct throughout a randomized field experiment at a bycicle messenger service in Zurich. Their findings show that the overall labor supply increases due to an increase in wages. In addition, they measure a higher elasticity of hours worked than the overall labor supply elasticity, meaning that workers reduce effort per hour in response to a wage increase.

The main inspiration for this thesis comes from the paper by Powell (2020). To the best of my knowledge, Powell is the first scholar that examines the impact transitory income shocks can have on household labor supply decisions in a more general way. In specific, he looks at the 2008 tax rebate program that was part of

[2]Heckman (1993) gives a detailed overview about the research development of labor supply theory.

the 2008 Economic Stimulus Act. Using monthly data from the panel of the Survey of Income and Program Participation (SIPP), Powell finds that households reduce their labor supply in response to the tax rebates issued in the course of the Economic Stimulus Payments (ESP) in 2008. His findings suggest that on average, each rebate dollar reduces labor earnings by 9 cents per month, with smaller but significant lagged effects. Households belonging to the second quartile of the earnings distribution show the largest reduction effects. Also, households indicating to have used the rebate payments for reducing their debt show strong evidence of labor reductions, which suggests that one of the main drivers of this effect might be liquidity constraints. It can thus be concluded that Powell not only establishes an initial link between transitory income shocks and labor supply decisions, but also measures a large heterogeneity in household responses. Overall, his estimations show that the $96 billion in stimulus payments resulted in a partial equilibrium reduction of short-term labor earnings by more than $26 billion, a reduction of more than 27 percent. Therefore, Powell's findings add an important component for both the reassessment of past and the assessment of future fiscal stimulus programs.

Heterogeneity in Labor Supply Heterogeneity in labor supply responses has, however, not been of major interest in the labor supply literature so far. Still, with the new and constantly advancing possibilities for heterogeneity measurements, such as quantile treatment estimations or tree-based methods in the field of machine learning, the barriers for these measurements are becoming less. This is reflected in the increasing number of scientific publications that investigate transitory shocks and their heterogeneous implications. As already mentioned in the introduction, some progress has been made in the area of consumer research. But also in the field of labour market research, there are first publications that measure heterogeneity in effects with quite some accuracy. For instance, Keane (2015) examines effects of labor income taxation in life-cycle models with human capital accumulation and finds that, among other things, human capital reduces responses of young workers to transitory tax changes and that labor supply elasticities depend on workers' age, with lower elasticities for younger and higher elasticities for older workers. Davis and Heller (2019) apply a machine learning approach to two randomized field experiments in order to measure heterogeneous employment impacts of summer jobs. In my thesis, I will rely on a machine learning algorithm to measure heterogeneous labor supply responses to transitory income shocks. Hence, I combine an effective approach to measure

heterogeneity with a research question analyzing labor supply responses to transitory income shocks. To the best of my knowledge, this thesis is one of the first scientific papers trying to address this question by means of a machine learning algorithm.

3 The Economic Stimulus Act of 2008

In this section I will provide some information on the overall economic situation in 2008 surrounding the tax rebates under consideration. Also, the motivation for introducing the stimulation package will be discussed. Besides, details regarding the design of the stimulus program are presented in order to establish a sufficient knowledge of the historical background that I am referring to in the further analysis.

The financial crisis of 2008 is generally seen as the worst economic crisis since the Great Depression, resulting both in the destruction of an estimated several trillion dollars on global stock markets and an increase in the number of unemployed people in OECD countries from a pre-crisis level of 34 million in 2008 to 42.1 million in 2010.[1] With the first signs of a burgeoning economic crisis, the American administration around president George W. Bush feared an upcoming recession of the U.S. economy and started considering how household income losses could be effectively mitigated and how the country's economy could be revived. These considerations led to the formulation of the so-called Economic Stimulus Act, which was published by the United States House of Congress and passed the U.S. House of Representatives on January 29, 2008. The U.S. Senate approved a slightly different version of the act on February 7, 2008. Thereupon, the Economic Stimulus Act came into force with the signature of president Bush on February 13, 2008.

The Economic Stimulus Act should allow taxpaying households to receive individual tax refunds from the state, which optimally would then be consumed by the households. The government's hope therefore was to provide a dynamic boost to the American economy through a household consumption channel. The act was projected to increase the total deficit by $152 billion, out of which $96 billion was attributable to the one-time Economic Stimulus Payments. According to Sahm et al. (2010), this total amount represented approximately 0.8% of personal income and 8.7% of federal tax payments in 2008.

The Economic Stimulus Payments consisted of three main parts: an individual income tax rebate for eligible recipients, sent in the month from April to July 2008,

[1] See OECD (2008) for more information.

and two provisions for businesses to encourage investments. Since my study concentrates on the effects of the individual tax rebates on labor supply decisions, I will from now on focus exclusively on the individual reliefs and generally describe them as 'tax rebates'. Table A.2 gives a brief overview of the payout dates for the rebate payments. Eligible recipients were individual or married taxpayers with an earned income of at least $3,000, which filed tax returns for either 2007 or 2008. The amount of the individual rebate payments was conditional on the individual's net income tax liability. The minimum rebate payments for low income households were $300 for individuals and $600 for married taxpayers, and it could not exceed $600 and $1,200, respectively. Additionally, a $300 rebate was granted for each child of an eligible individual or married household taxpayer. Rebate reductions took place at a rate of 5% of the adjusted gross incomes over $75,000 (or $150,000 for married taxpayers respectively). Therefore, as Broda and Parker (2014) point out, rebates were below $300 or even zero both for households with low and high enough incomes. In total, stimulus payments amounted to almost $100 billion and were paid to approximately 120 million individuals in the U.S. in 2008.

On the one hand, the timing of the rebate payments depended on whether the tax filer provided a bank routing number on the 2007 tax return. On the other hand, it depended on the last two digits of the Social Security number (SSN). Starting in late April 2008, electronic fund transfers (EFTs) for filers that provided their bank routing number were completed on May 16, whereas individuals that received their payment checks by mail potentially had to wait for their rebate until July 11. However, if the recipient submitted her 2007 tax return later in the year, it could happen that payments were made later in 2008 (Powell, 2020). Since the receipt of payments depended on households' arbitrarily assigned social security numbers, the timing itself was randomised. Consequently, the payments can be considered a natural experiment. This characteristic will be of great importance for the implementation of the machine learning approach, which will be discussed in more detail in chapter 4.

4 Empirical Approach

This chapter mainly serves to explain the theoretical background of the empirical strategy I follow throughout my analysis. First, in section 4.1, I will describe the origin and the handling of the data. Second, section 4.2 introduces the reader to the two main theoretical concepts which my estimation relies on. I start with a brief explanation of uplift modeling and go more into detail when describing the idea behind random forest algorithms.

4.1 Data

Survey of Income and Program Participation The American Survey of Income and Program Participation (SIPP), a national panel data set based on household observations, serves as the data source for this Master's thesis. According to the United States Census Bureau, the SIPP collects information of household's labor income, labor force participation and social program participation as well as other, more general demographic characteristics, such as education, health insurance, and child care.[1] Since its launch in 1983, the overall purpose of the survey is to mirror a quite realistic picture of the income and program participation situation of American households, with a special focus on income distribution and measurement of general well-being. It is considered the most extensive information source available to study the national well-being over time. Historically, the United States Census Bureau initiated the survey in order to tackle three main objectives. First, annual and sub-annual income dynamics should be evaluated. Second, a tracking of movements into and out of the government's transfer programs should be possible. And third, the measurement of family and social context of individuals and households should be facilitated.

Technically, the SIPP questions are centered around core modules including questions on labor force, program participation and income. The survey can be supplemented by topic-specific questions, which can be added to the list of questions due to current events or other reasons. Households within the sample are interviewed in a regular four-monthly cycle. This four month period is

[1] More information on the history, content and use of the SIPP can be found here.

called a wave. Within every interview, households must state their monthly labor earnings as well as participation status and demographic characteristics for each month passed. As a result, the SIPP tracks income dynamics and labor status for participating households for every single month in a wave.[2]

Since the Economic Stimulus Payments took place in 2008, the analysis of the Master's thesis will be based on the 2008 panel of the SIPP. Here, the focus will be on the first two waves of the 2008 panel. Due to the Economic Stimulus Payments, which began in April 2008 and ended in July 2008, the first two waves include special questions about the ESPs. In specific, households were asked to report the size as well as the timing of the additional rebates they received in the course of the stimulus payments. Since interviews in the first two waves were staggered across households, wave 1 and 2 contain 8 months of data for each household, starting from May 2008 and ending in March 2009. According to Powell (2020), the total stimulus payments indicated in the survey account for 96% of the overall payments that were reported by the Department of the Treasury. Thus, the impact measurable with the help of the survey data can be considered representative for all American households receiving stimulus payments.

Data Handling The raw SIPP data provided online by Powell builds the data base for this thesis. As mentioned in subsection 4.1, I concentrate my analysis on data in the first two waves of the 2008 SIPP panel data set only. Consequently, I exclude information from other waves as well as households with negative or missing labor income data in the beginning. Households with a reported monthly labor income of 0 aren't paid attention to since I am investigating labor income responses at the intensive margin only. Furthermore, I limit my sample on households whose head is between 25 and 60 years old and I exclude cohabiting, but non-married households because of the possibility of two rebate receipts.[3] I do not allow households to change their marital states within the observation months, since this would dynamically change the numbers of single and married households in the data.[4]. Households are restricted to a maximum family size of six. Also, the lower and upper five percent of households in terms of reported labor income are excluded. Moreover, I do not allow labor income to

[2]With minor exceptions, topcoding is applied to observations with $\geq$\$12,500 of monthly labor earnings.

[3]Out of the roughly 23,000 households included in the beginning, only 3 report multiple rebate payments.

[4]For the sake of simplicity, I set each households' initial value as the default value for the remaining seven observation months.

vary extremely between months.[5] 9,981 households in waves 1 and 2 are affected by these manipulations.[6]

My cleaned data sample includes variables for the measurement of the monthly reported labor income, household size, marital status, age of the households' head as well as variables indicating whether the households' head is self employed or gets paid on an hourly basis. Further, variables for year, month and wave indication are included. Also, several characteristics of the rebate payments are covered by different variables, such as information about rebate receipt, rebate sizes and use of payments.[7]

Following Powell (2020), I consider household labor income as the target variable, and not the sum of individual income in households, since I assume that a married couple living together in one household coordinates its labor decisions and that, consequently, a separate measurement would lead to inaccuracies. Therefore, the main analysis will use the reported family income of an household as the target feature. From a more practical point of view, this also makes sense since the majority of fiscal policy programs is based on household behavior rather than on individual behavior of each member of a household.

Table A.1 shows some income statistics for the remaining data set, listed by month of payment receipt and split into single and married households. Rebate payments in the months of April and May were received electronically primarily. In total, 12,863 households with eight monthly observations each are included in the sample for the main analysis, out of which 4,769 households have a single and 8,094 a married head. 85.5% of single households and 91.6% of married households in the sample received a rebate payment.[8] The average rebate amount was \$596 for single households and \$1,138 for married households. Income statistics are quite uniformly distributed over the months of rebate receipt.

Out of the 12,863 households that I observe eight times each, 5,574 households receive a rebate in one of the observation months. 5,913 households receive their rebate before their first month of observation. Thus, I observe 1,376 households that do not receive a rebate payment in any month. I include these households for fitting the algorithms. As a result, my estimation model will have 5,574 observations in the so-called treatment group and 97,330 observations in the control

[5] I exclude households who report monthly labor incomes of greater or equal to 1.5 times their mean labor income as well as smaller or equal to 0.67 times their mean labor income within the observation period of 8 months. However, it turns out that my model would yield similar estimation patterns if strongly varying observations were included.

[6] Apart from minor differences, data cleaning is based on the logic of Powell (2020).

[7] A detailed description of the variables included can be found in the online documentation.

[8] These percentage numbers are quite high also because I excluded zero labor income households that normally did not receive a rebate payment.

group. After having trained the algorithms, I predict both treatment group labor income and control group labor income for rebate months and months afterwards only. Consequently, the number of final predictions is smaller than the total size of my initial data sample. I further explain the idea of splitting the sample into two groups for estimating treatment effects in the following section 4.2.

Additionally, the SIPP provides information on weekly hours worked, employment status and reasons for work absences. Unfortunately, this data is not aggregated to the household level and only available for individuals. Therefore, I conduct a second, somewhat less detailed analysis based on individuals. The structure of the second data set is identical to the previous one, but I define the individual labor income as the new target variable and include variables for determining working hours and absences in the prediction model. A comparison of the two data sets shows that 87.6% of the individuals in the second sample are represented in the first data set in aggregated form, measured as members of a household. Therefore, I emphasize that the two data samples do not necessarily represent observations of the same individuals. Being aware of this, I will still conduct this analysis in order to gain more insights about the heterogeneous reactions of individuals to transitory income shocks. Moreover, since data on working hours within reference months is available, the second sample on individual responses allows to derive a rough estimate of working hour changes rather than labor income changes. Estimations for these changes will be presented in chapter 5.7.

4.2 Empirical Strategy

After having discussed the SIPP and the data cleaning, the focus of the remaining part of this chapter will be on theoretical considerations with respect to both the uplift modeling and the random forest algorithm. The following explanations mainly serve to create a basic understanding of the statistical methods used in the empirical analysis.

Uplift Modeling

Recall that the ultimate goal of my thesis is to measure labor supply changes as a reaction to transitory income shocks in the form of tax rebates. In particular, this means that I want to measure the behavior of a household when it gets a rebate, and how it behaves when it does not get one. As mentioned in chapter 3, one can consider the rebate payments a natural experiment due to their

random timing. This fact can be exploited by splitting the overall sample into a treatment group, which includes observation months in which rebate payments are received, and a control group, which consists of all other months in which households do not receive a payment. By using the random forest for predicting labor income levels conditional on different features for both groups and taking the difference between the two predictions, I am able to determine the impact of rebate payments on labor income. In machine learning theory, this approach is called uplift modeling (Jaroszewicz, 2017). Uplift modeling is often used for statistical analyses in the medical or marketing sector, where there is a particularly high interest in measuring treatment effects. Generally, there are two ways to tackle the implementation of uplift models. The first approach, which I refer to as the one-model or direct approach, involves a direct computation of differences. The second approach is the so-called *two-model approach*. In contrast to the direct method, two models are built and predictions are made parallel to each other. Taking the differences between both predictions for each observation after the separate estimations are made yields the treatment effect one is usually interested in. For my application, the two model approach is an intuitive extension of the random forest model, in the sense that two forests are built parallel to each other, one for the treatment group predictions and one for the control group predictions. Therefore, I will rely on the two-model approach when computing the labor income differences. Calculating labor income differences in my case will be of the form

$$\widehat{LID}_i = \underbrace{\hat{f}_{LI}(x_i|T)}_{\textit{treatment group prediction}} - \underbrace{\hat{f}_{LI}(x_i|C)}_{\textit{control group prediction}} \tag{4.1}$$

where i runs over all months with and after rebate payment, LID stands for labor income difference, $\hat{f}_{LI}$ represents the labor income predictions, T indicates treatment and C stands for control. Intuitively, one observation can either be treated or not treated at the same time. Hence, only one outcome $\hat{f}_{LI}$ can be observed. I circumvent this by predicting labor income levels for both states. Therefore, I construct two data samples in the first step - a treatment group consisting of household observations in months in which rebates are received, and a control group with all other observations. Second, I train both random forest models with their respective training set. Last, I predict labor income for observations from rebate months or months after rebate payments only. Put differently, I use the entire data set to fit the prediction model but I will exclude months prior to rebate payments for the final predictions. As a result, this approach allows me to treat each household observation as if it was both treated and untreated in the same month. The differences form a vector that displays the labor income change.

Random Forest

This section aims to introduce the random forest algorithm. It contains explanations on the characteristics of the random forest, the idea and theory of regression trees as well as of random forests and a brief discussion of partial dependence plots as a measure for heterogeneity in treatment effects.

Introduction to Random Forests Although tree-based methods have only become more famous in recent years, the idea of using decision trees to classify data has existed since the 1960s. To the best of my knowledge, Morgan and Sonquist (1963) were one of the first that described the construction of decision trees. 20 years later, Breiman et al. (1984) covered the use of trees as an instrument for data analysis in a mathematical framework, and therefore can be considered the first to really link classification and regression trees (CART) to analyzing large data sets.

Nowadays, tree-based methods increasingly complement the statistical methodology of economists. In particular, trees have the ability to identify nonlinear relationships between different variables. This is an enormous advantage over linear models, which by definition only measure linear interactions in data sets. For instance, the usually applied linear regression design only measures the average effect in data samples. In order to measure heterogeneous effects anyway, scholars usually resort to subgroup analysis. Unfortunately, also subgroups normally vary in many different characteristics, which makes it impossible to pin down the specific effect of every driver. In contrast, the use of random forest algorithms allows to understand the effect of each variable individually. By varying one variable at a time while holding all other variables constant, the marginal effect of each variable is measurable without distortions (Gulyas and Pytka, 2019). Since overall, random forests seem to be very appealing for estimating heterogeneous treatment effects, I choose this algorithm for my estimation.

Regression Tree I begin the theoretical consideration of random forests with a small preliminary remark on simpler tree-based methods. The motivation for this is intuitive. I will explain the concept of one regression tree first, and then extend it to several parallel trees, which build a so-called forest.

Regression trees represent a break from classical methodology, such as generalized linear models (GLMs). Unlike GLMs, regression trees rely on recursive partitioning, a technique that, based on defined criteria, splits a data set into smaller subsets until a stopping rule is reached. One can think of this as a creation of data bins, each of which results in a different prediction value. Regression

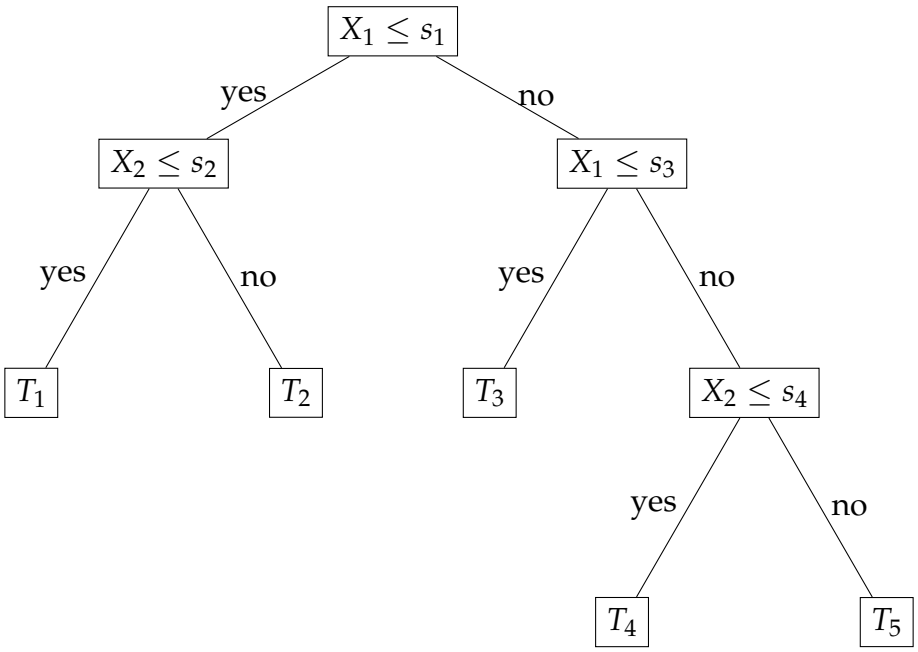


FIGURE 4.1: Illustration of an exemplary tree

trees are non-parametric, computer-intensive and were promoted by bigger but less structured data sets in the recent past. Figure 4.1 illustrates the method using a textbook example from Efron and Hastie (2016). Using an exemplary data set with a dependent variable Y and two independent variables X_1 and X_2, the tree uses X_1 to split the initial data at the first node. Observations with X_1 values smaller or equal than s_1 are put into the left branch, whereas X_1 values larger than s_1 are put into the right one. This procedure is repeated in each newly emerging node with perhaps different variables and splitting criteria under the premise to provide the highest prediction accuracy until previously defined stopping criteria are reached.

In this example, this results in five terminal nodes $T_1, T_2, ..., T_5$, for each of which a separate prediction is created. This is true for categorical and regression trees. I will use regression trees in my analysis, so I want to specify this approach for regression trees in more detail.[9] Imagine a data set with observations (x_i, y_i) for $i = 1, 2, 3, ..., N$ and $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$. Suppose that the algorithm partitions into M regions. Let $\hat{c}_m$ be a constant response for $m = 1, ..., M$ such that

$$\hat{f}(x) = \sum_{m=1}^{M} \hat{c}_m \mathbb{1}(x \in T_m). \tag{4.2}$$

Usually, the splitting point at the nodes is chosen such that the squared sum of

[9]The mathematical explanations below strongly follow the representation in Hastie et al. (2009).

residuals $\sum_i (y_i - \hat{f}(x_i))^2$ is minimized. Let N_m be the number of observations in T_m. It follows that

$$\hat{c}_m = \frac{1}{N_m} \sum_{i \in T_m} y_i. \tag{4.3}$$

Finding the best split is done through a greedy approach. In the first step, the algorithm uses the splitting variable j and the split point s to define

$$T_L(j,s) = \{X | X_j \leq s\} \qquad \text{and} \qquad T_R(j,s) = \{X | X_j > s\}\,, \tag{4.4}$$

where T_L is the left branch and T_R is the right branch leaving a particular node in the tree. In order to derive the optimal values for j and s, the algorithm solves

$$min_{j,s}[\, min_{\hat{c}_L} \sum_{i \in T_L(j,s)} (y_i - \hat{c}_L)^2 \quad + \quad min_{\hat{c}_R} \sum_{i \in T_R(j,s)} (y_i - \hat{c}_R)^2]\,. \tag{4.5}$$

As mentioned above, the inner minimization with respect to $\hat{c}_L$ and $\hat{c}_R$ is solved by

$$\hat{c}_L = \frac{1}{N_L} \sum_{i \in T_L} y_i \quad \text{and} \quad \hat{c}_R = \frac{1}{N_R} \sum_{i \in T_R} y_i, \tag{4.6}$$

respectively. After having calculated 4.6 and solved 4.5 by computing the split point s for every j, the algorithm derives an optimal pair (j, s) for the splitting node. This procedure is repeated until the predefined stopping criterion is met. While regression trees are easy to interpret, they often tend to overestimate relations in the data. This so-called *overfitting* problem arises when the complexity of the estimation model is so high that it mistakenly confuses noise with structural patterns in the data. When applied to new test data of the same population, the performance of overfitted models is usually poor. Overfitting can be controlled for by using bootstrap aggregation (*bagging*), which is one of the key elements of the random forest algorithm I will use for my analysis.

Random Forest Random forests were introduced by Breiman (2001). In general, the idea behind a random forest is to reduce estimation variance by averaging. This is achieved by growing many regression trees to randomized versions of the training data and averaging them (Efron and Hastie, 2016). The algorithm itself makes use of bootstrap aggregation (*bagging*), which I will briefly describe in the following.[10]

Bootstrapping was first introduced by Efron (1979) and, in its simplest way, refers to B rounds of drawing N times from a sample $x = (x_1, x_2, x_3, ..., x_N)$ with

[10] Mathematical notions follow Hastie et al. (2009) once again.

replacement, resulting in B bootstrap-samples $x^{*b} = (x_1^{*b}, x_2^{*b}, ..., x_N^{*b})$, $b = 1, ..., B$. These B bootstrap-samples can then be used for different sorts of computations, e.g. approximating the a priori unknown probability distribution F of the original sample x or estimating prediction values and errors. As an extension of bootstrapping, bagging averages predictions based on bootstrap samples and thereby reduces its variance. Recall the exemplary data set (x_i, y_i), $i = 1, ..., N$ with prediction value $\hat{f}(x)$ from above. For each bootstrap sample (x^{*b}, y^{*b}), with $x^{*b} = (x_1^{*b}, x_2^{*b}, ..., x_N^{*b})$, $y^{*b} = (y_1^{*b}, y_2^{*b}, ..., y_N^{*b})$ and $b = 1, ..., B$, the B prediction values will be of the form $\hat{f}^{*b}(x)$. Averaging over all bootstrap samples gives the bagging estimate

$$\hat{f}_{bagging}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x). \tag{4.7}$$

Since trees are generally noisy, report a low bias if grown deeply and capture complex interactions in the data, they are an ideal candidate for bagging (Hastie et al., 2009). Still, bagged trees have the same bias as each individual bootstrap tree. As a result, bagging cannot reduce a forest's bias. However, there is a good reason for its use since it can effectively lower the variance of random forest predictions.[11] From

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{4.8}$$

which describes the variance of a bootstrapped sample of size B, it can be seen that for increasing B, the second term converges to zero, and the variance of the bootstrap average depends solely on the pairwise correlation of the trees ρ and the variance of the trees σ^2. Hence, reducing ρ while keeping σ^2 on a low level will lead to variance reductions. In random forests, this is achieved by randomly selecting only a fraction $l \leq k$ of the feature variables before each split. A possible form of the random forest algorithm is sketched on the next page.[12]

As shown earlier, single trees are quite interpretable - graphically displaying them allows to understand the splitting variable j and the split point s. With large forests, the interpretation is not so easy anymore (Varian, 2014). Especially, since splitting variables are randomly selected, they differ substantially between trees, therefore making comparisons more difficult. However, there are possibilities to get a glimpse into this random forest black box. Below, I explain the idea of partial dependence plots. Another option to better understand random forest predictions are feature importances. I will give some theoretical explanations on the determination of importances when discussing them for my predictions in section 5.6.

[11] A mathematical explanation can be found in Appendix C.

[12] Algorithmic representation based on Hastie et al. (2009).

Algorithm Random Forest

1. **For** $b = 1$ **to** B**:**
   (a) Draw a bootstrap sample (x^{*b}, y^{*b}) of size N from the training data.
   (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the stopping criterion is met.
      i. Select randomly l variables from the k overall variables.
      ii. Choose the minimizing solution (j, s) according to (4.5) and split the single node into two daughter nodes.
2. **Gather the trees** $\{T_b\}_1^B$**.**
3. **Make predictions** $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Partial Dependence Plots Quantifying the influence of a feature on the target variable is crucial for the interpretation of the results. When using random forests for estimations, this can be done using so-called partial dependence plots (Friedman, 2001). In general, partial dependence plots are illustrations that plot the random forest prediction $\hat{f}(x)$ against the feature of interest x. For the sake of dimensionality, it is useful to select a small subset of variables only. To make the idea a bit more vivid, I assume that my data set consists of N observations and has k explanatory variables. Since I am only interested in the partial dependence of my target variable from l features, I choose a subset $x^s = (x^1, ..., x^l)$ from the original set of variables $x = (x_1, ..., x_k)$. Note that the l features in x^s do not have to be the first l variables. Define x^c such that $x = x^s \cup x^c$. In general, the prediction of the target variable $\hat{f}(x)$ is a function of x. Therefore, it applies in particular:

$$\hat{f}(x) = \hat{f}(x^s, x^c). \tag{4.9}$$

By marginalizing over the distribution of the features in x^c, we can derive the partial dependence function for regressions:[13]

$$\hat{f}_{x^s}(x^s) = \mathbb{E}_{x^c}\left[\hat{f}(x^s, x^c)\right] = \int \hat{f}(x^s, x^c)\, d\mathbb{P}(x^c). \tag{4.10}$$

[13]The original derivation by Friedman can be found in Appendix C.

In the sample, what I will do is estimating the partial dependence function by averaging over the training data, which will look like this:

$$\hat{f}_{x^s}(x^s) = \frac{1}{N}\sum_{i=1}^{N}\hat{f}(x^s, x_i^c). \tag{4.11}$$

Thus, the predictions $\hat{f}_{x^s}(x^s)$ display the estimate for a given value of feature x^s, holding the other features x^c constant for all observations $i = 1, ..., N$.

By using an example, I will explain the formula above a bit more in detail. Suppose that I am interested in quantifying the influence of the variable *famsize*, which indicates the number of persons in a household for a given observation month, on the labor income. Consequently, x^s only contains *famsize* and x^c consists of all the other features used for the estimation. I form a grid g, which contains the different values of *famsize*: $g = \{1, 2, 3, ..., n\}$. Assuming that each observation in the data has $j \in g$ persons in its family, I estimate the labor income for each observation, while keeping the other features constant. For the final prediction, I take the mean value. I repeat this procedure for all values $j \in g$. It becomes quite obvious that this method might require an enormous computational effort for large grids g.

One caveat of partial dependence calculations in this way is the assumption of uncorrelation between the features in x^s and the features in x^c. If this assumption is violated, this leads to inaccurate estimates because then, averaging over x^c would consider unrealistic combinations of feature values that would distort the final prediction. Although being very intuitive, partial dependence plots neither incorporate the feature distribution nor do they show heterogeneous responses for one particular value of x^s.[14]

There are standard Python packages for the computation of partial dependence plots. However, in my particular application these already implemented algorithms do not seem to effectively help answering the research question properly. Therefore, I build a tailor-made algorithm that calculates directly the difference in labor income between treatment and control predictions for each value of each feature.

[14]To partially offset these limitations, I add boxplots, which describe the distribution of the feature in the respective observation period.

5 Main Results

In this chapter I will present the results of the random forest labor income predictions using uplift modeling. Moreover, I will discuss possible mechanisms and explanations for the changes in labor income. In doing so, I will rely on calculations of feature importances and partial dependencies.

5.1 Overall Treatment Effects

I start with focusing on the results of the analysis taking into account households' reactions in rebate receiving months and all months after. For my analysis, I use a random forest with 10,000 parallel grown trees and a maximum depth of 5. I do not include household weights when training the algorithm. Hence, for computing the optimal splitting variables and splitting points, each household in the sample is weighted identically. I include household weights for the extrapolation in section 5.4, however.

First and foremost, I emphasize that the stability of the estimates with varying parameterization is not necessarily given. This clearly aggravates the interpretation of the results. As a result, my interpretations are limited and should be treated with caution. Further, a look at the distribution of absolute labor income responses in Figure 5.1 shows that absolute estimates of households' labor income changes due to rebate receipts as well as changes relative to the reported income level are widely scattered. I find large outliers especially in the positive area of labor income changes in both absolute and relative terms.[1] This suggests that the distribution of estimated labor income changes is skewed to the right, with mean values being to the right of median changes.

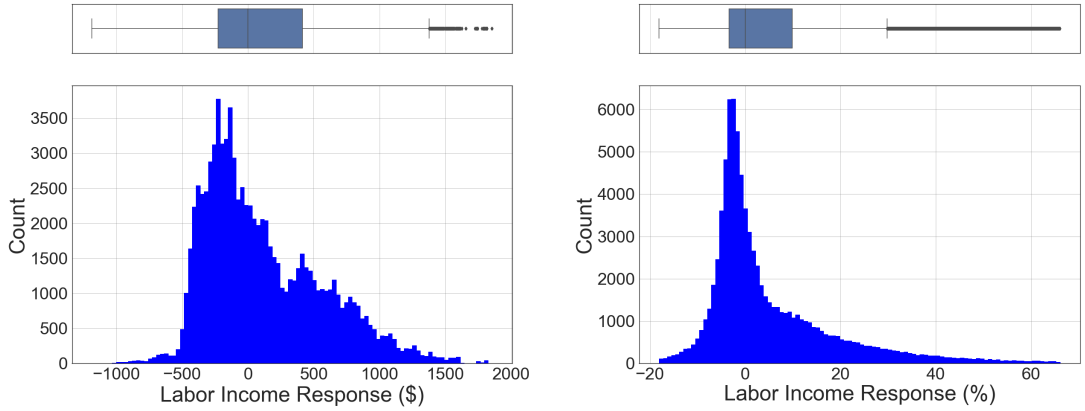
This observation is confirmed by the results in Table 5.1. While the mean monthly changes are well into positive territory for all three measurement methods, the results of the median household are slightly negative, but close to zero. In fact, the median household reduces his monthly labor income by 16 cents per observation month. Since households differ significantly in their reported monthly labor income, absolute income changes are limited in terms of validity. Setting

[1] Labor income responses relative to the rebate size show a similar distribution.

TABLE 5.1: Labor Income Responses in All Observation Months

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	115.064	-0.164	450.372	-1,185.158	-226.937	414.271	1,856.987
(2)	0.055	-0.000	0.159	-0.698	-0.034	0.102	1.533
(3)	0.516	-0.000	12.555	-318.696	-0.205	0.668	1,074.513

Estimations are based on B = 10,000 trees. The results are based on 87,522 estimations from 11,487 households.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.



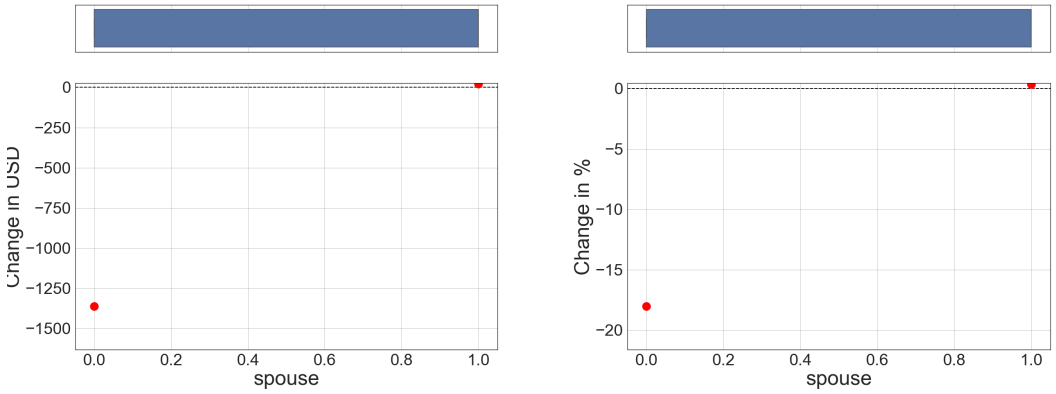
(A) Absolute Labor Income Response (B) Percentual Labor Income Response

FIGURE 5.1: Distribution of Labor Income Responses in All Observation Months

Note: Lower and upper 1% are excluded from the graphical illustration in (B). The shown boxplots illustrate the distribution of labor income responses in all observation months.

the predicted absolute change in labor income in relation to the reported monthly labor income and the rebate size allows to better interpret the estimated labor income responses. Results for these relative expressions are shown in the second and third row in Table 5.1, respectively.[2] Relative changes are smaller than -0.1% both compared to the reported labor income and the rebate size received. Hence, taking into account all observation months in the data set, I find no long-term change in labor income due to rebate payments for the median household. However, results look different when considering mean responses of households. The calculated mean values are distorted by strong positive outliers. On average, households increase their labor income by about $115 per month due to the rebate payments. The marginal change in terms of rebate size is even more extreme.

[2]Labor income response estimates of the three models show strong positive rank correlations. I report correlations in Table A.3 in the appendix.



(A) Median Absolute Responses (B) Median Percentage Responses

FIGURE 5.2: Partial Dependence Plot for *spouse* in All Observation Months

Note: Results based on 87,522 estimations from 11,487 households in rebate months and months after. The shown boxplots illustrate the distribution of labor income responses for different values of the feature *spouse* in all observation months. The observations contain 4,076 single and 7,411 married households.

When considering all months, an additional dollar in rebate payments leads to a 51.6% increase in monthly labor income, on average. The high variance of the estimates strongly limits the mean as the primary indicator of response measurements. For this reason, I will primarily rely on the responses of the median household as the reference point for the interpretation of my results.

At the beginning of this chapter, I have already briefly stated that the high variance in estimation suggests individual household responses to differ substantially. As discussed, Figure 5.1 gives a first idea of this. However, I am able to investigate the heterogeneity more precisely using partial dependence plots (PDPs). More specifically, partial dependence plots allow to get a deeper understanding of how a single feature in the data sample affects labor income differences between treatment and control predictions. Figure 5.2 illustrates this effect using the example of the feature *spouse*, which measures the marital status of a household.[3] What the estimates show is that single households reduce labor income strongly, whereas married households do not seem to adjust their labor income after a rebate payment. Responses are similar for the relative measurement in Figure 5.2b. These strong differences in responses within each prediction variable can also be observed with other features in the data set. I find that households of size one reduce labor income stronger than households with two or more members both in absolute and percentage terms. This result is in line with the observation that non-married households are more prone to income

[3]Figure 5.2 displays median responses because I want to commit to one measure.

TABLE 5.2: Labor Income Responses in Rebate Months

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	-64.961	-89.553	422.648	-1,185.158	-327.733	190.065	1,486.616
(2)	-0.027	-0.017	0.110	-0.698	-0.075	0.034	0.533
(3)	-0.181	-0.091	9.857	-280.158	-0.473	0.189	588.582

Estimations are based on B = 10,000 trees. The results are based on 5,574 estimations from 5,574 households.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

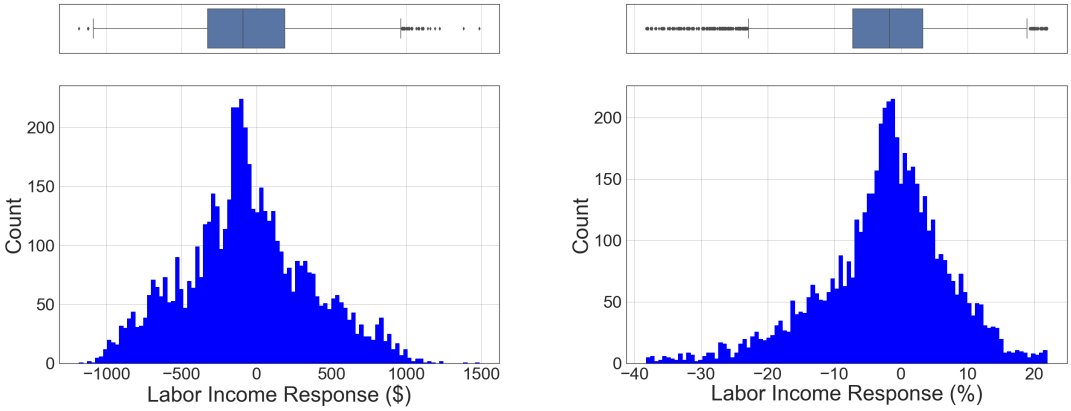
reductions than married couples. Also, households with at least one person being paid on an hourly basis reduce labor income strongly, whereas this reduction cannot be observed for households with workers whose salary is not based on hours worked. Further, labor income reductions seem to be driven by younger people. My model predicts that households with heads younger than 40 years lower labor income when all observations months are considered. In Appendix B, I present partial dependence plots for multiple features.[4]

5.2 Treatment Effects in Rebate Months

In this section, I will focus on months in which rebate payments occur only. Measured changes during these months quantify the immediate response of households to the randomised payments. The results are illustrated in Table 5.2. The median household reduces its labor income by around $89.6 in months of rebate receipts. In relation to the rebates paid, I find that the median household reduces its labor income by about 9 cents for every dollar of rebate paid. Despite the high variance of the results, it appears that households generally react more strongly to rebates in the rebate month itself. In particular, the more negative mean response and also the strong negative mean percentage reactions in terms of labor income and rebate size indicate that there are fewer positive outliers in the rebate months. This assumption is supported by the somewhat lower variance of all three estimation models compared to all observation months before. It therefore seems plausible to conclude that in rebate months, households reduce labor income stronger and behave less diversely than in the remaining months.

Also in rebate months a high degree of heterogeneity in responses can be observed. Compared to the responses to rebate payments in all months, I observe

[4]I show plots for median absolute responses only. Plots for percentage point and marginal changes can be found in the Supplementary Appendix online.



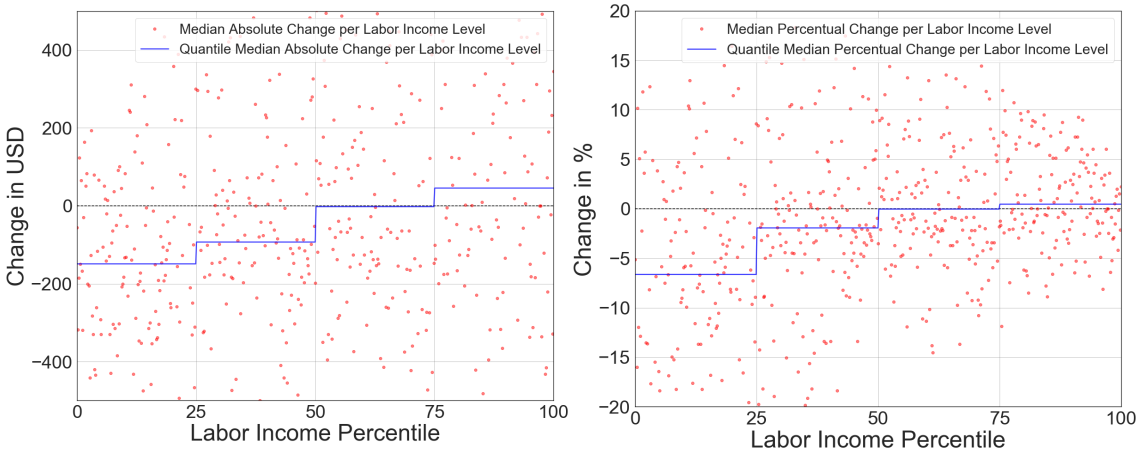
(A) Absolute Labor Income Response (B) Percentual Labor Income Response

FIGURE 5.3: Distribution of Labor Income Responses in Rebate Months

Note: Lower and upper 1% are excluded from the graphical illustration in (B). The shown boxplots illustrate the distribution of labor income responses in rebate months.

stronger immediate labor income reductions in both absolute and percentage terms (Figure 5.3). Especially the distribution of labor income changes relative to reported monthly labor income levels changes significantly in rebate months only (Figure 5.3b). This is in line with the results from Table 5.2, since both the mean and the median labor income are reduced more strongly in rebate months than in all observation months combined. Moreover, Figure 5.4 illustrates the differences in responses for different labor income groups in rebate months only. In particular, it shows median absolute and median percentage labor income responses for each quartile of the labor income distribution. I am aware that the estimates are very noisy and that mean and median values are not really reliable. Also, this classification by income level contradicts my analysis using partial dependence plots for estimating the "pure" impact of a feature. To avoid hidden characteristics, partial dependence plots are definitely preferable to subgroup analysis, and they are the main reason why I decided to use a machine learning algorithm to tackle my research question. That being said, I still rely on this kind of subgroup analysis at this point since I define labor income as the target variable and therefore I am unable to construct partial dependence plots for labor income levels as an estimation feature. I must take this detour in order to be able to at least roughly estimate the impact of different labor income levels.[5] In doing so, I find

[5]In fact, lagged labor income as a feature could have enabled PDP analyses for labor income levels. However, this means losing the first observation of each household, which is equivalent to a loss of 2,784 household observations in the rebate month or a reduction of treatment group observations by 49.9%.



(A) Median Absolute Response

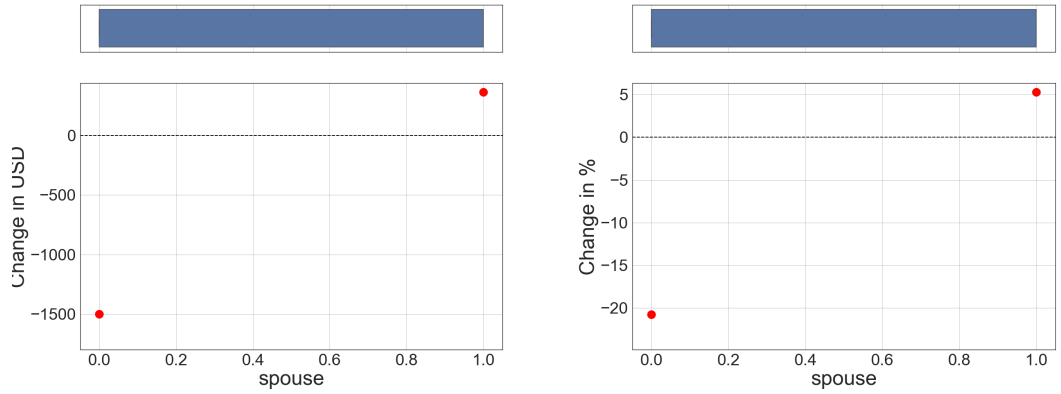
(B) Median Percentage Response

FIGURE 5.4: Labor Income Responses for Different Labor Income Quantiles in Rebate Months

Note: Results based on 2,972 estimations for 2,972 different labor income levels in rebate months only.

that the first, second and third labor income quartiles show negative responses (Subfigure 5.4a). In percentage points of households' reported monthly labor income, as Subfigure 5.4b illustrates, similar patterns can be observed. Except the upper quartile, whose labor income response in percentage points is equal to 0.46%, the vast majority of households reduces labor income. The lowest quartile of the labor income distribution responds in the strongest way, both in absolute and percentage terms with labor income changes of -$149.02 and -6.64%, respectively. For each rebate dollar received, the lowest quartile reduces labor income by around 23%, while the second quartile's reduction is 8.55%. Again, only the richest fraction of households reacts with an income increase (+4.28%) for each rebate dollar received (Figure B.2).

Heterogeneous responses can also be found when looking at different prediction features. In order to display changes between all months and the months in which the rebates were paid, I will show responses of the *spouse* feature again. Compared to Figure 5.2 the results in Figure 5.5 are based on estimates in rebate months only. The distribution of heterogeneity between the different realizations of the *spouse* feature remains. It is noticeable, however, that responses in rebate months are stronger in both absolute and relative terms. This is true not only for the feature *spouse*, but also for the majority of other features in the estimation model. For instance, in contrast to considering all months, my model estimates negative responses for all ages in rebate months alone. Moreover, I observe strong heterogeneities in the labor income change conditional on the size of the rebate. Based on my estimates, the majority of rebate receiving households reduces labor income in absolute terms. However, rebates of around $1,100 or more result in



(A) Median Absolute Responses (B) Median Percentage Responses

FIGURE 5.5: Partial Dependence Plot for *spouse* in Rebate Months

Note: Results based on 5,574 estimations from 5,574 households in rebate months only. The shown boxplots illustrate the distribution of labor income responses for different values of the feature *spouse* in rebate months. The observations contain 2,000 single and 3,574 married households.

positive labor income changes. When comparing households according to their stated use of the rebate, I find evidence of liquidity constraints as a potential driver of labor income reductions, since households paying back their debt with parts of the rebate reduce labor income significantly. Surprisingly, also households indicating to have used rebates for consumption spending reduce labor income in rebate months. In Figure B.4, I illustrate labor income responses for a variety of features, including rebate size and stated use of the rebates.

5.3 Lagged Treatment Effects

In this section I will focus on households' reactions in the two months after rebate payment combined. In doing so, I want to obtain a more realistic picture of the effects since it is quite conceivable that significant household responses to labor income could also be observed in the months after rebate payment. Table 5.3 illustrates the lagged estimation results for the selected period. I find a less negative median absolute estimate of -$38.1 for the two months combined. Relative median estimates are negative but closer to zero than in the rebate month as well. It appears that the median household's response is strongest in the rebate month itself, and it slowly decreases the longer the payment dates back. Tables A.6 and A.7 back up this finding. Again, mean estimations seem to be influenced by outliers at the right end of the distribution. This is true for all three measurements of responses. Therefore, I report a positive mean absolute labor income response of about $19.7 as well as quite large relative numbers for (2) and (3). Tables A.4 and

TABLE 5.3: Labor Income Responses in Lagged Months

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	19.694	-38.095	336.918	-775.573	-248.531	269.891	1,605.327
(2)	0.026	-0.006	0.105	-0.386	-0.039	0.076	1.091
(3)	0.269	-0.037	8.918	-318.696	-0.225	0.439	559.510

The first two months following rebate payments are considered in the estimation. Estimations are based on B = 10,000 trees. The results are based on 18,415 estimations from 10,202 households.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

A.5 show labor income responses in both lags separately. I find more negative median responses in the first lag than in the second month after rebate receipt. While the median household reduces its monthly labor income by $40.65 in the month after rebate, it does so by $35.38 in the subsequent month. Similar patterns can be observed when measuring household responses as fractions of labor income levels and rebate size. Once again, mean estimates seem to be distorted by outliers.

Similar to what I have found before, also household responses in both lagged months contain a large amount of heterogeneity. Partial dependence plots, which consider reactions in the first two months after rebate receipt, can be found in Figure B.5.

5.4 Overall Partial Equilibrium Effect

In addition to estimating the monthly labor income responses for each household in the data sample, I am also interested in transferring these estimates into a measurement of the overall macroeconomic impact of the rebates in the months I considered before. However, since I am only estimating direct effects on labor income changes and do not further exploit their consequences on other macroeconomic variables, I limit my interpretation of the monthly labor income changes to partial equilibrium effects. I follow the computation of Powell (2020) and build $\tau \in \{1, ..., 100\}$ labor income quantiles. The idea is to compute the accumulated labor income responses in three steps. First, I calculate the accumulated labor income response that my data sample represents. For each of the 100 labor income quantiles, I do this by multiplying the sum of marginal effects for different lags with factors that take into account the number of households, the household

weights and the rebate sizes in the particular quantile.[6] Having done this, I integrate over all 100 quantiles to get the accumulated labor income response that my data sample actually represents.[7] Second, I scale back the accumulated labor income response to a \$1 rebate by dividing it by the amount of rebates my data sample represents. Last, multiplying this marginal response for a \$1 rebate receipt by the size of the stimulus package yields the aggregate partial equilibrium effect in the months considered. A mathematical formula calculating the partial equilibrium effect could look like this:

$$PEE_t = \int_\tau \left[N(\tau) \cdot \frac{\sum_{i=1}^{N(\tau)} weight_i}{N(\tau)} \cdot \frac{\sum_{i=1}^{N(\tau)} rebate_i}{N(\tau)} \cdot \widehat{marg}_t(\tau) \right] \mathrm{d}\tau \cdot \frac{1}{RS} \cdot ESP\,,$$

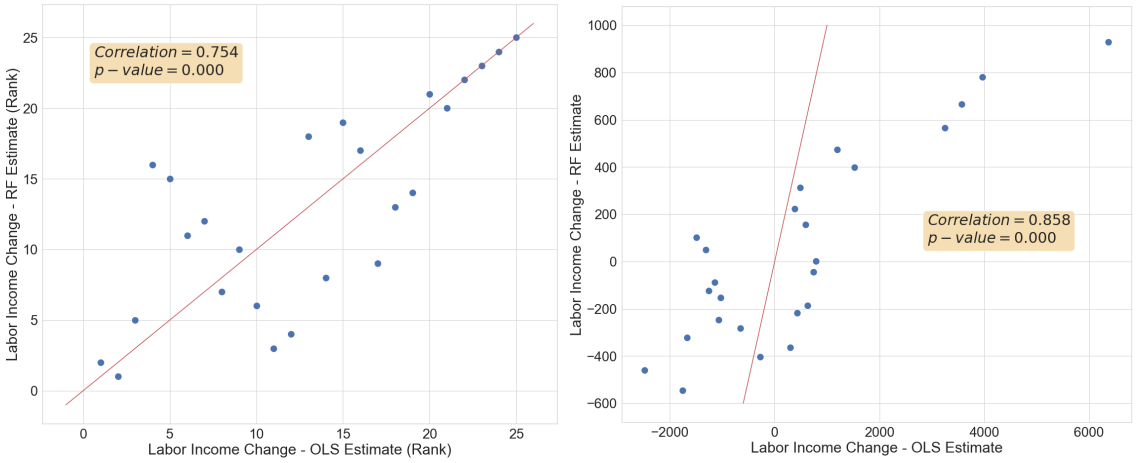
where PEE_t is the partial equilibrium effect for time period t, τ is the income quantile, $N(\tau)$ is the number of observations in quantile τ, $weight_i$ represents observation i's weight[8], $rebate_i$ is observation i's rebate size and $\widehat{marg}_t(\tau)$ is the sum of mean responses relative to rebate size for quantile τ and time period t. Further, RS is the product of rebate size per household in the sample and each household's weight, and ESP is equal to the \$96 billion of total payments.

For $t = \{0, 1, 2\}$, so the rebate month itself and the two consecutive lags, I find a partial equilibrium effect of -\$12.2 billion. This corresponds to a reduction of 12.7% of the total size of the stimulus package. Not including the third lag in my calculation, my estimates show a partial equilibrium effect less than half as large as Powell's \$26.7 billion partial equilibrium reduction of national labor earnings. However, my extrapolation should be treated with caution. It seems that instead of yielding more accurate and plausible results, the weighting factors included combined with the high variance of marginal response estimates bias the outcome for each month alone. Nonetheless, I include this extrapolated result in the analysis because it helps in getting a better understanding of the magnitude of the effects. When measured against the total rebate sum of \$96 billion, a reduction of \$12.2 billion in the first three months is much more tangible than just the mean or median effect per household on a monthly basis.

[6] I exclude marginal responses smaller than -150% and larger than 150% for each quantile.

[7] I choose this approach because when aggregating quantiles should be weighted with both their average rebate and their average household weights. Otherwise, I would weight every quantile identically, which would bias the accumulation.

[8] Observation i's weight indicates the number of American households that is represented by observation i in the specific observation month.



(A) Rank Scatter Plot RF against OLS

(B) Bin Scatter Plot RF against OLS

FIGURE 5.6: Random Forest Estimation Accuracy

Note: Estimation accuracy measured as correlation between random forest estimates and OLS estimates. A high correlation can be associated with a high accuracy. P-values measure the statistical significance of the correlation. The red lines display 45 degree lines for orientation.

5.5 Accuracy Measurement

Measuring the accuracy of my estimates is not as straightforward as it seems at first glance. With commonly used measures of accuracy (e.g., the MAPE) the idea is to compare the estimations with corresponding values in the data and then measure the percentage deviation. In my application, however, this is not possible since I estimate a treatment effect for which there is no benchmark value in the data. Circumventing this, I apply an approach by Gulyas and Pytka (2019), who group individuals into bins according to their estimated treatment responses. Then, they estimate treatment responses using OLS and compare them with the random forest results. Hence, they get a rank correlation showing whether the random forest estimates are in line with OLS estimation, regardless of the magnitude of estimation values. I will conduct the same analysis, with the only difference that I group households instead of individuals into 25 bins. Results are presented in Figure 5.6. Despite having some discrepancies especially in the medium rank size, I observe a highly significant positive rank coefficient of 0.754, particularly driven by same rank grouping at the lower and upper end of the estimations (Figure 5.6a). The correlation of the actual estimated values in the right panel is somewhat higher at 0.858 and also statistically significant. Similar to Gulyas and Pytka (2019), I also find that the OLS estimation overfits towards positive outliers in particular and therefore suggests a larger degree of heterogeneity in responses.

5.6 Driving Features of the Predictions

One of the main problems with machine learning algorithms is that they are sometimes described as black boxes. The main reason for this is that at first glance, it often is not obvious how predictions or classifications were made. Which features play an important role? Which ones are more predictive, and which ones may be completely disregarded? For all those questions the simple estimation result does not provide answers. However, in the practical application of these algorithms, it often is of great importance to identify the crucial channels that lead to the estimate.

There are ways to answer the questions above. A popular approach is the calculation of the so-called feature importances, which measure how important individual features are for predicting values. The actual measurement of importances can be done in multiple ways, and should therefore be considered carefully before assessing the results of individual features. The RandomForestRegressor algorithm from the scikit-learn package in Python already has a predefined command that allows the user to measure the importances quite easily. Results for my estimation are shown in Figure 5.7.

Since I calculate my final estimates of labor income responses using the two-model uplifting approach, I get two feature importance values for each feature. The red bars indicate the importance for the predictions of the control group random forest (not receiving rebate payments) and the blue bars measure importances for the predictions of the treatment group random forest (receiving rebate payments). Since features that include information on the rebate size and the use of rebates are only used for the treatment group random forest predictions, I do not observe their importances for the control group random forest. Feature importances for all features can be found in Table A.10 in Appendix A. I find that by far, the binary feature *spouse*, which differentiates single households from married ones, is the most important feature in both the treatment and the control group random forest. *flex* (measures the number of hourly paid workers in a household) and *famsize* (number of persons in a household) are the second and third most important features according to the impurity based default measurement. Moreover, it seems that the size of the rebate (*rebate uplifting*) is a driving channel for treatment predictions in my model. Still, at first glance it is quite unclear on what basis these importances were computed. In fact, the predefined command uses the standard *mean decrease in impurity* mechanism (also called *gini* mechanism) that measures the total decrease in node impurity (variance), scaled by the probability of reaching this node for each feature (Breiman et al., 1984). Certainly,

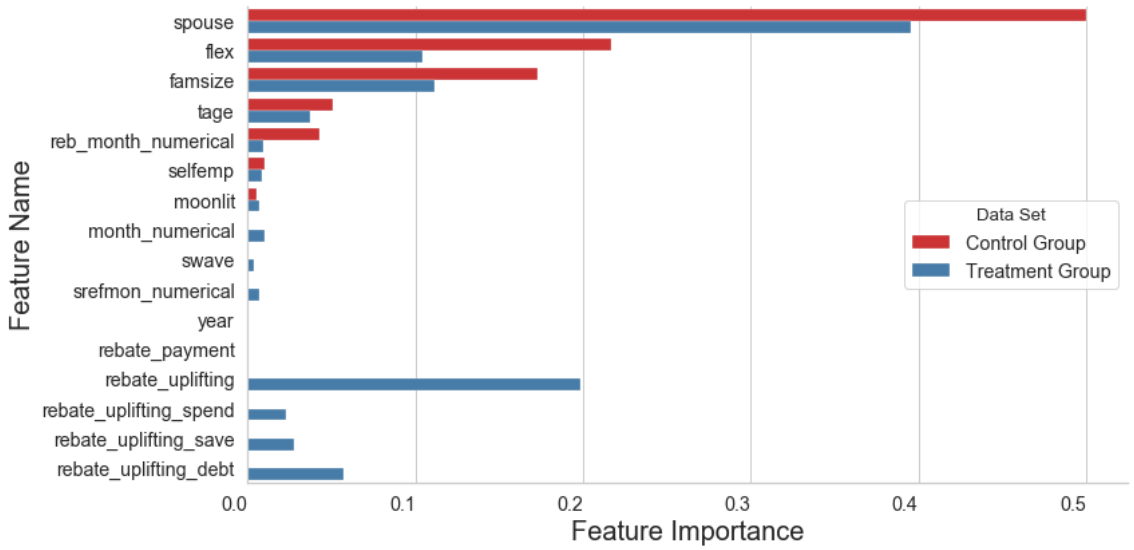


FIGURE 5.7: Mean Decrease in Impurity Feature Importance

Note: Computation based on the default feature importance measure already implemented in the RandomForestRegressor algorithm.

this is a fast and intuitive way to measure the feature importance.[9] However, there are certain caveats using this approach. Strobl et al. (2007) show that feature importances are distorted because they generally overstate the importance of high-cardinality features. Additionally, importances are computed based on the training data. In the worst case, when the algorithm overfits extremely, they do not tell anything about the drivers for the final predictions. Therefore, it seems reasonable to compare the impurity based importances with a second method, which is based on feature permutations.

The idea of the *permutation feature importance* is to randomly permute the values of the feature of interest, and compare the prediction error of the permuted model with the original one. First, the prediction accuracy of each tree is recorded. Second, one predicts again, while this time randomly permuting the variable of interest in the sample. If the model error of the permuted model is significantly higher than the one from the original model, the predictions seem to rely on this feature. Therefore, a higher importance should be attached to it. In contrast to the default computation of feature importances, the permutation importances are not scaled in order to sum up to 1. In fact, the relative differences between features display their predictive power, whereas the absolute values are not of great importance. When applying this approach to my data, I still attach the highest importances to the variables *spouse*, *flex* and *famsize* (Figure 5.8). Note that the

[9]The predefined command's running time is around 2 seconds for my example. Using permutation importances, it takes 22 minutes already. Last, when dropping each feature in order to measure its importance, the running time increases to 127 minutes approximately.

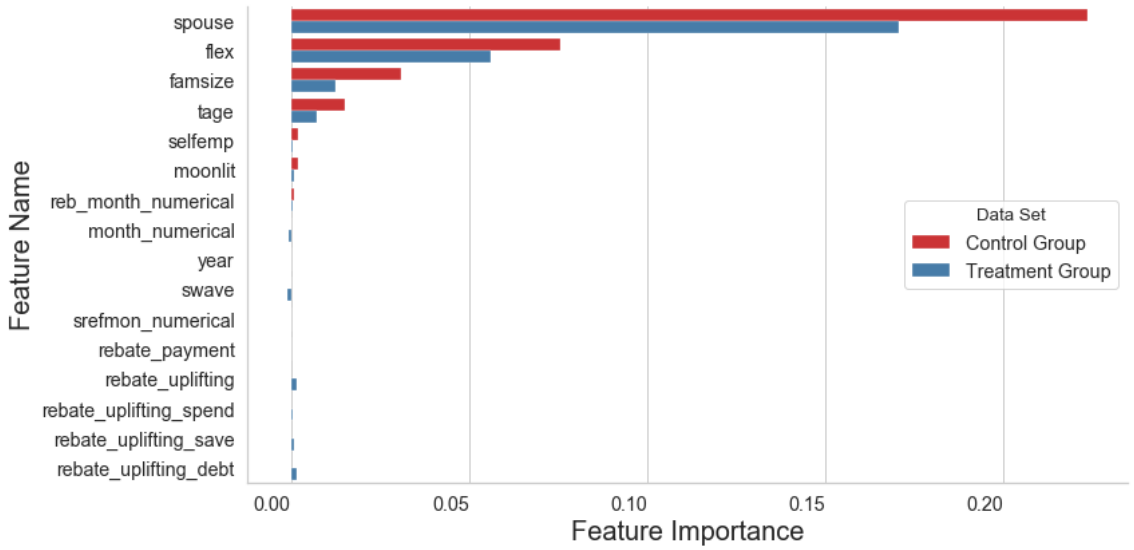


FIGURE 5.8: Permutation based Feature Importance

Note: Computation based on the Permutation Importance command implemented in the eli5.sklearn package.

size of the rebate does not play a crucial role for the treatment predictions anymore. One explanation might be that, in fact, due to the high cardinality of *rebate uplifting* the impurity based calculation overestimates its real influence for predictions. However, it is quite surprising that the size of the rebate only plays a minor role, since one would assume that households change income based on the size of the payment. One of this approach's main advantages is that the permutation feature importances are computed on the final predictions, whereas the default computation uses the training set only. Further, high cardinality features are not systemically overestimated. Also, the approach does not require the retraining of the model every time a different feature is assessed. This is extremely time-saving, as the third approach below shows. One caveat, however, is the calculation's sensitivity to correlation in the features. If features are highly correlated, and one of them is randomly reshuffled, this leads to highly unrealistic combinations of features that are then used for predictions. However, this issue does not really occur in my analysis since my features are generally quite uncorrelated. I show Pearson and Spearman correlation matrices of my estimation features in the Supplementary Appendix online.

The third, most direct approach to measure the importance of one feature involves excluding the variable of interest from the random forest algorithm and comparing its performance in terms of variance reduction with the situation in which the variable is included. Although being the most "honest" way to compute one feature's importance, it is extremely time-consuming since it requires repeated model training. Results can be seen in Figure B.1 in the appendix. The

TABLE 5.4: Individual-level Working Hour Responses in Rebate Months

	mean	**median**
Change (in working hours)	-3.998	-2.260
Change (relative to total working hours)	-0.018	-0.013
Change (per $1,000 rebate payment)	-9.055	-2.577

Estimations are based on N = 10,000 trees. The results are based on 7,791 estimations from 7,791 individuals.

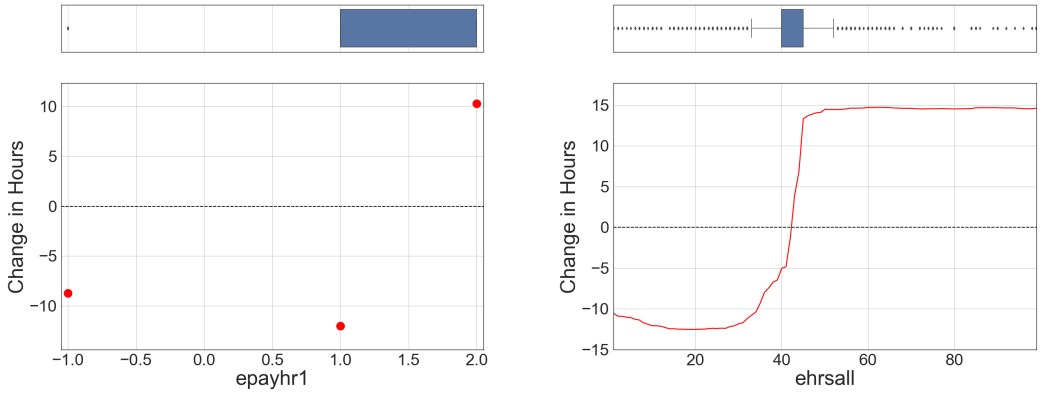
rank of the features is quite the same, while absolute values differ. Note that negative feature importances occur, which means that dropping the feature from the analysis would actually increase the predictive power of the model.

5.7 Individual Working Hour Responses

Unfortunately, detailed information on hours worked and reasons for work reductions are provided at the individual level only. Hence, in order to be able to draw conclusions about the actual change in labor supply measured by hours, I apply my estimation model to the data set based on individual observations. Data cleaning is identical to the household-level data sample.

Table A.8 presents labor income changes on the individual level for rebate months only.[10] The median individual reduces labor income in the month of rebate receipt by $47.67, or 1.3% of her reported labor income. Every rebate dollar leads to a reduction of 5.4 cents. Mean estimates are more negative for all three calculations. Using the reported monthly labor income and the number of monthly working hours, I calculate an average hourly wage for each individual's observations. This allows me to express the estimated labor income difference in terms of hours worked. Changes in labor supply measured in hours worked are shown in Table 5.4. I find a median individual labor supply reduction of 2.26 hours and an average labor supply reduction of around 4 hours in rebate months. In terms of total monthly hours worked, individuals reduce labor supply by 1.8% on average, with a slightly smaller negative median response. Further, the median response suggests households to reduce labor supply by around 2.6 hours in rebate months alone per rebate payment of $1,000. Again, average effects are more negative. Aggregate household-level responses both measured in absolute hours and hours per $1,000 rebate payment are shown in Table A.9. I measure significant heterogeneity in working hour responses. Figure 5.9 illustrates that workers who are paid by the hour and people usually working around 20 hours

[10] I focus on rebate months only to keep the analysis on the individual level brief.



(A) Median Absolute Responses for Workers Paid by the Hour

(B) Median Absolute Responses for Usual Hours Worked per Week

FIGURE 5.9: Heterogeneity in Working Hour Responses

Note: Results based on 5,574 estimations from 5,574 households in rebate months only. (A): 1.0 indicates paid by the hour, 2.0 not paid by the hour, -1.0 for no information. The shown boxplots illustrate the distribution of labor hour responses for different values of the features *epayhr1* and *ehrsall* in rebate months.

per week reduce their working hours strongly. This suggests part-time employees in particular to be one of the main drivers of labor hour reductions. In addition, individuals appear to reduce their working hours largely on a voluntary basis, as the strongest reduction in hours is observed for "personal days" as a reason for reducing work.[11]

[11] More partial dependence plots showing heterogeneous working hour responses on the individual level can be found in Figure B.6.

6 Conclusion

This thesis studies the impact of tax rebate payments on labor supply decisions of households. To do so, I build an empirical model that uses two random forest algorithms to make monthly labor income predictions which I use to compute treatment effects for each household in the sample. My approach allows me to document heterogeneity in labor income responses. Although my model shows a rather high variance in estimation results, I find an immediate median labor income reduction of 9 cents for each rebate dollar received. Further, I measure decreasing median reductions for consecutive months after payment. Labor income responses are driven by young and single households as well as households with hourly workers. Also, liquidity constraints might be a crucial factor since households indicating to use the rebate payment to pay back debt seem to lower labor income in particular. While short-term labor income responses are negative, I do not estimate long-term median changes. Intuitively, this is difficult to explain, as it implies that households increase their labor supply in the long run to potentially counteract the short-term reduction.

When studying the impact in rebate months on the individual level, I find similar responses. For a $1,000 rebate, individuals reduce working hours by 9 hours per month, on average. Estimated median hour changes are still negative, but smaller. In particular, these reductions seem to be voluntary and mostly driven by individuals with hourly wages and part-time workers.

The estimated short-term labor supply reductions suggest that the proven consumption stimulus due to fiscal stimulation packages is only one side of the coin. I show that in the short run, the overall stimulation effect might be dampened by the reduction in the supply of labor. Although it seems that households steer against this reduction in the long run, my analysis shows that households react sensitively and heterogeneously to such transitory income shocks. Thus, I provide evidence that lump sum payments themselves could be an effective instrument to affect households' short-term labor supply decisions.

A Appendix A

TABLE A.1: Main Analysis Data Set Summary Statistics

	Full Sample		Rebate Receiving Sample		Rebate Receiving in April/May		Rebate Receiving in August–December	
	Single	Married	Single	Married	Single	Married	Single	Married
Family size	1.79	3.38	1.79	3.37	1.78	3.39	1.71	3.41
Received rebate (%)	85.50	91.60	100	100	100	100	100	100
Average rebate size in $ (conditional on receipt)	596	1,138	596	1,138	606	1,155	574	1,123
No monthly labor earnings (%)	0	0	0	0	0	0	0	0
25th percentile monthly labor earnings	2,400	4,150	2,400	4,200	2,524	4,344.75	2,333	4,167
Median monthly labor earnings	3,500	6,234	3,456	6,233	3,584	6,317	3,724	6,342
75th percentile monthly labor earnings	5,167	8,717	4,933	8,617	5,083	8,750	5,083	8,845.75
Number of households	4,769	8,094	4,076	7,411	1,509	2,888	291	512

Illustration based on Powell (2020).

Note: "Single" and "Married" refer to marital status in the household's first observation in the data. This marital status can change over the 8 months of data, although such changes are rare. Each household is represented in the data for 8 months.

TABLE A.2: Timing of Economic Stimulus Payments in 2008

Last two digits of SSN	Electronic transfer made by	Check in mail by
00-20	May 2	
21-75	May 9	
76-99	May 16	
00-09		May 16
10-18		May 23
19-25		May 30
26-38		June 6
39-51		June 13
52-63		June 20
64-75		June 27
76-87		July 4
88-99		July 11

Illustration based on Powell (2020).
Note: The form of payment depends on whether the taxpayer has indicated her bank identification code on the tax return in 2007. SSN stands for Social Security Number. Later receipt of payment possible.

TABLE A.3: Correlation Table for Labor Income Response Estimates

	Pearson Correlation	Pearson p-value	Rank Correlation	Rank p-value
(1) - (2)	0.868	0.000	0.955	0.000
(1) - (3)	0.095	0.000	0.963	0.000
(2) - (3)	0.086	0.000	0.937	0.000

Estimations are based on B = 10,000 trees. The results are based on 87,522 estimations from 11,487 households. Rank correlation measured as Spearman correlation.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

TABLE A.4: Labor Income Responses in First Month after Rebate Payment

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	8.614	-40.649	329.256	-757.778	-252.663	255.138	1420.642
(2)	0.023	-0.007	0.101	-0.370	-0.040	0.072	0.797
(3)	0.204	-0.040	7.703	-318.696	-0.232	0.419	466.933

The first month following rebate payments is considered in the estimation. Estimations are based on B = 10,000 trees. The results are based on 8,235 estimations from 8,235 households.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

TABLE A.5: Labor Income Responses in Second Month after Rebate Payment

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	28.656	-35.377	342.729	-775.573	-245.247	287.362	1,605.327
(2)	0.029	-0.006	0.109	-0.386	-0.038	0.078	1.091
(3)	0.322	-0.034	9.792	-276.791	-0.219	0.458	559.510

The second month following rebate payments is considered in the estimation. Estimations are based on B = 10,000 trees. The results are based on 10,180 estimations from 10,180 households.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

TABLE A.6: Labor Income Responses in Third Month after Rebate Payment

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	60.999	-27.623	376.186	-745.792	-240.882	340.787	1,856.987
(2)	0.039	-0.004	0.126	-0.386	-0.036	0.089	1.368
(3)	0.393	-0.025	10.246	-283.977	-0.211	0.567	625.092

The third month following rebate payments is considered in the estimation. Estimations are based on B = 10,000 trees. The results are based on 11,235 estimations from 11,235 households.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

TABLE A.7: Labor Income Responses in Fourth Month after Rebate Payment

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	-111.560	1.618	433.671	-739.038	-229.666	430.377	1,788.248
(2)	0.056	0.000	0.153	-0.396	-0.034	0.109	1.398
(3)	0.533	0.001	12.434	-258.475	-0.209	0.722	783.989

The fourth month following rebate payments is considered in the estimation. Estimations are based on B = 10,000 trees. The results are based on 11,385 estimations from 11,385 households.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

TABLE A.8: Individual-level Labor Income Responses in Rebate Months

	mean	**med**	**std**	**min**	**25%**	**75%**	**max**
(1)	-64.870	-47.669	166.388	-601.919	-177.042	48.538	1,191.755
(2)	-0.018	-0.013	0.064	-0.567	-0.047	0.014	0.439
(3)	-0.120	-0.054	3.993	-160.572	-0.197	0.045	269.136

Estimations are based on B = 10,000 trees. The results are based on 7,791 estimations from 7,791 individuals.
(1): Labor Income Difference in USD.
(2): Difference relative to Reported Labor Income.
(3): Difference relative to Rebate Size.

TABLE A.9: Household-level Working Hours Responses in Rebate Months

	mean	**median**
Change (in working hours)	-5.057	-3.106
Change (per $1,000 rebate payment)	-11.454	-3.788

Estimations are based on B = 10,000 trees. The results are based on 7,791 estimations from 7,791 individuals, which are aggregated to 6,159 households. Due to computational issues I do not show household-level changes relative to total monthly working hours.

TABLE A.10: Feature Importances

Feature	Mean Decrease in Impurity		Permutation based Importances		Dropping Columns	
	Control Group	Treatment Group	Control Group	Treatment Group	Control Group	Treatment Group
spouse	0.500	0.396	0.224	0.171	0.093	0.105
flex	0.216	0.105	0.076	0.056	0.041	0.051
famsize	0.172	0.112	0.031	0.012	0.004	0.021
tage	0.051	0.037	0.015	0.007	0.007	0.018
reb month numerical	0.042	0.010	0.001	0.001	-0.003	0.012
selfemp	0.010	0.008	0.002	0.001	-0.002	0.011
moonlit	0.006	0.007	0.002	0.001	-0.002	0.011
month numerical	0.001	0.010	0.000	-0.001	-0.004	0.011
swave	0.000	0.004	0.000	-0.001	-0.004	0.010
srefmon numerical	0.000	0.007	0.000	0.000	-0.004	0.010
year	0.000	0.000	0.000	0.000	-0.004	0.011
rebate payment	0.000	0.000	0.000	0.000	-0.004	0.011
rebate uplifting	-	0.198	-	0.001	-	0.010
rebate uplifting spend	-	0.023	-	0.000	-	0.010
rebate uplifting save	-	0.027	-	0.001	-	0.010
rebate uplifting debt	-	0.057	-	0.002	-	0.010

Note: Missing values occur if a feature is not included in the model. Negative values indicate an accuracy improvement when dropping the feature. Permutation based importances are calculated using the *eli5* algorithm.

B Appendix B

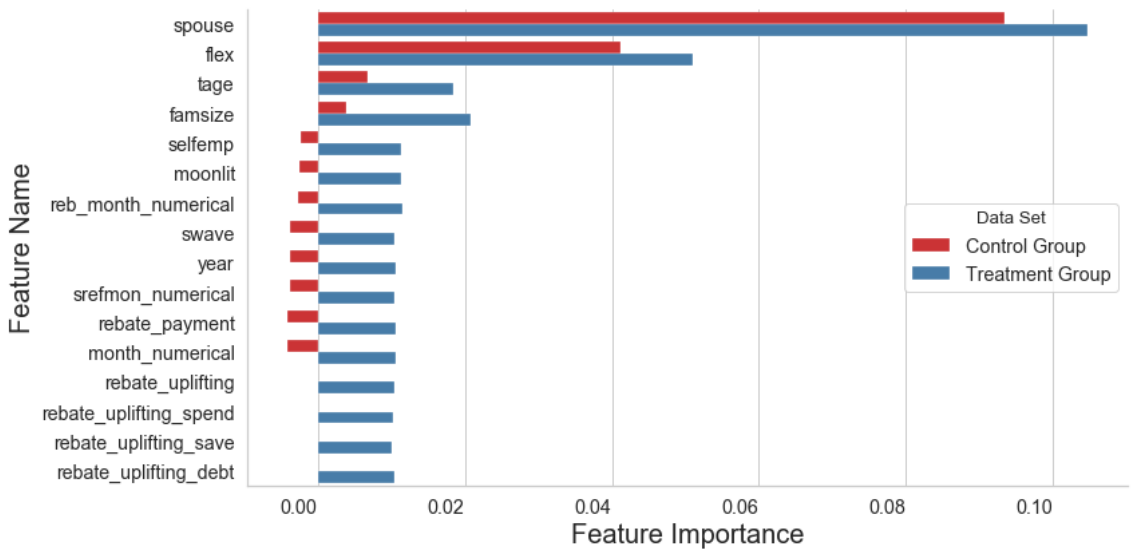


FIGURE B.1: Dropping Columns for Measuring Feature Importance

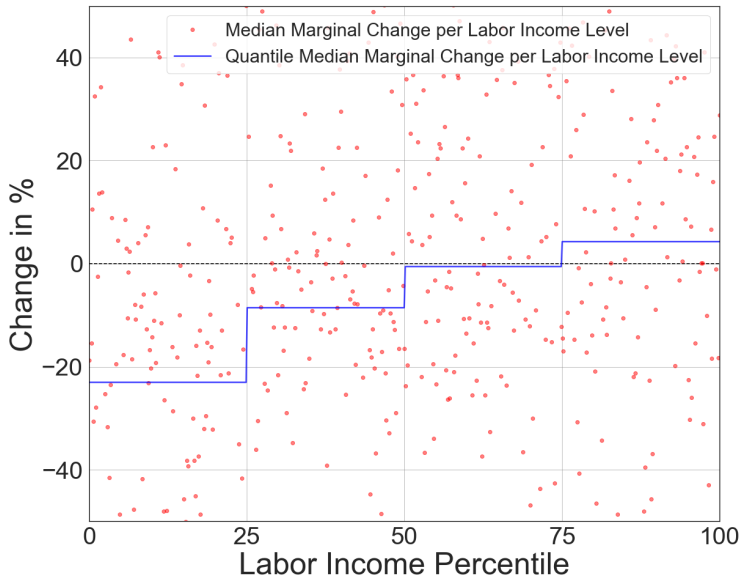
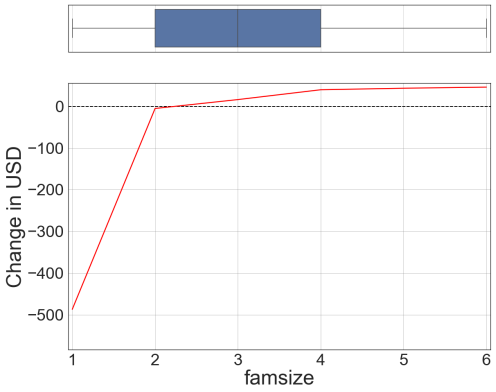
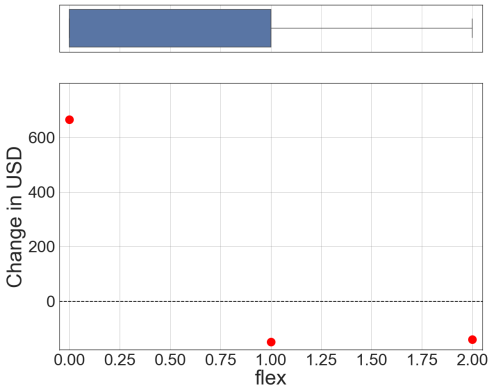


FIGURE B.2: Median Marginal Labor Income Response for Different Labor Income Quantiles in Rebate Months

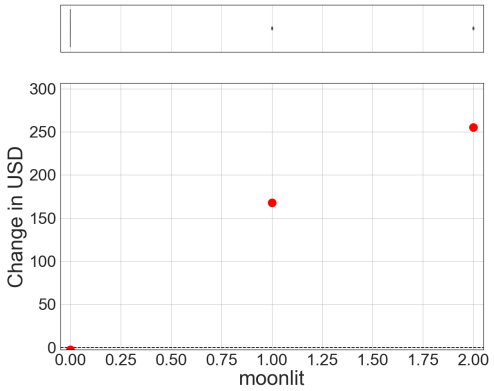
Note: Results based on 2,972 estimations for 2,972 different labor income levels in rebate months only.



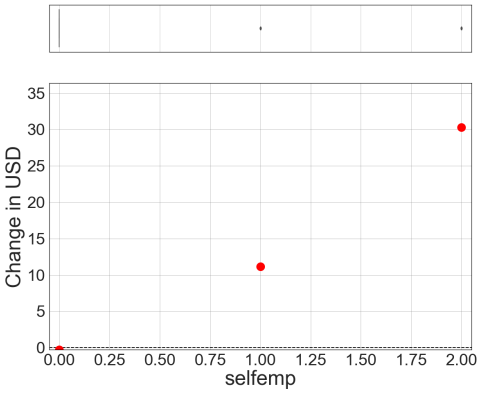
(A) family size



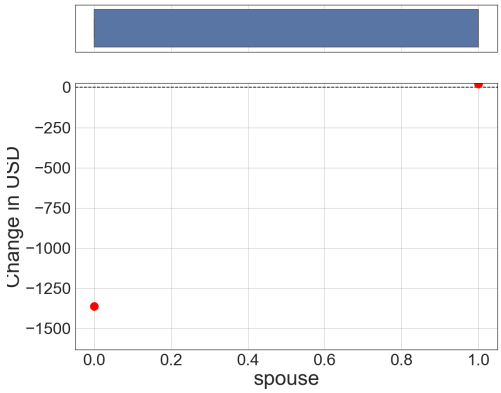
(B) number of flexible workers



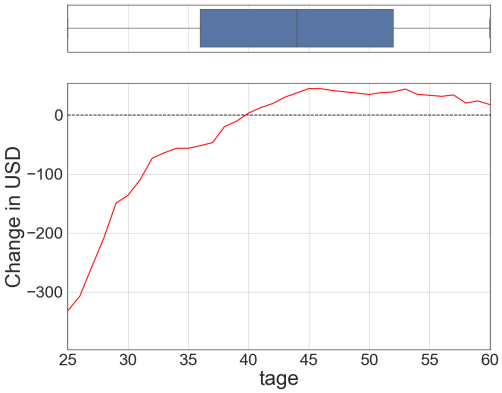
(C) extra income from non regular work



(D) self employed workers

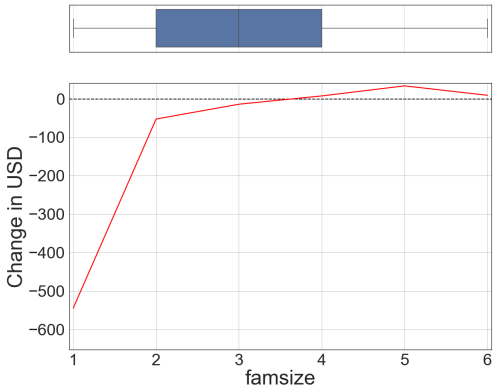


(E) married

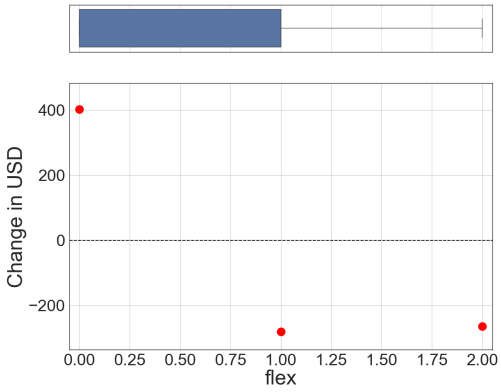


(F) age

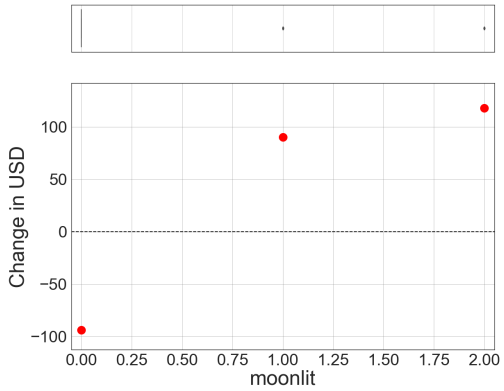
FIGURE B.3: Median Absolute Responses in All Observation Months



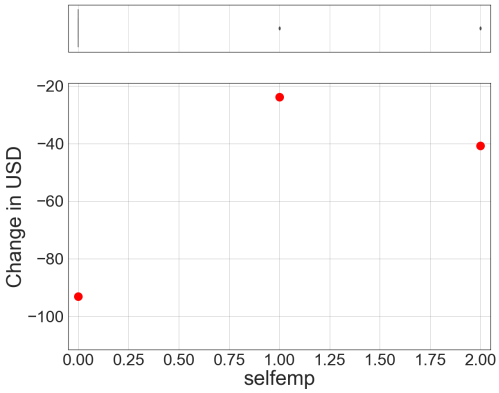
(A) family size



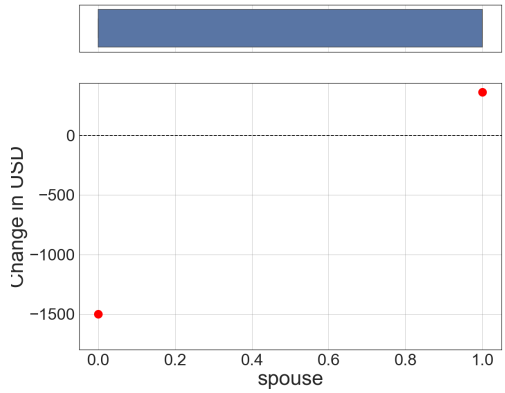
(B) number of flexible workers



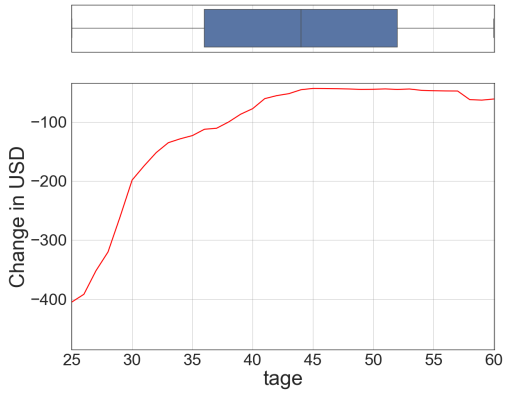
(C) extra income from non regular work



(D) self employed workers



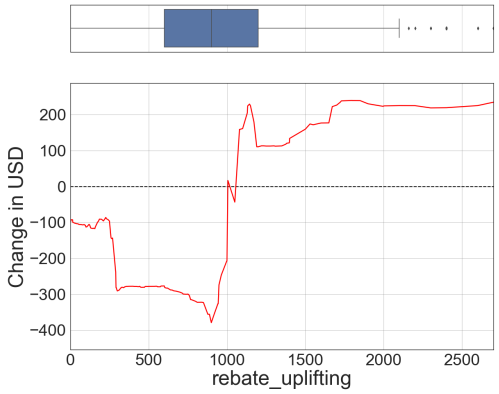
(E) married



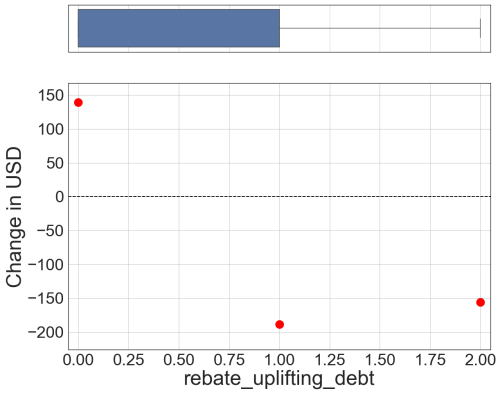
(F) age

FIGURE B.4: Median Absolute Responses in Rebate Months

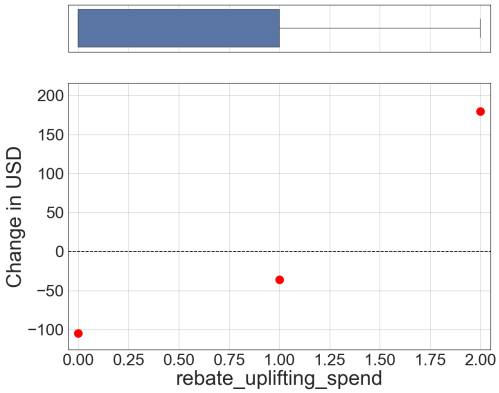
Note: For rebate months, I include partial dependence plots for features describing rebate characteristics, such as the rebate size and the reported use of the rebate payment.



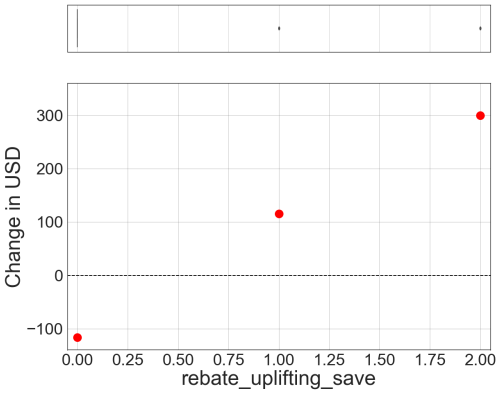
(G) rebate size



(H) rebate used to pay off debt



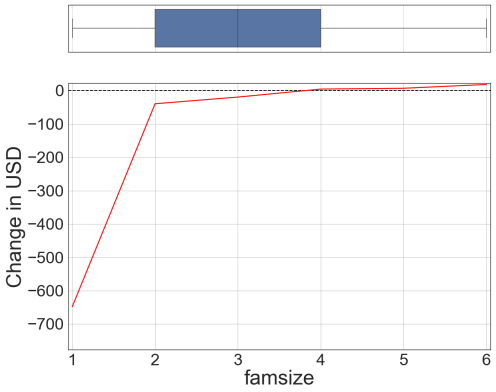
(I) rebate used to consume



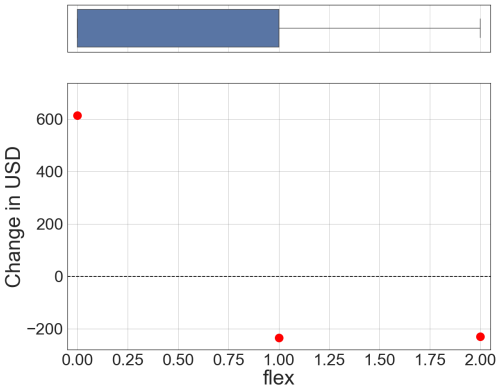
(J) rebate used to save

FIGURE B.4: Median Absolute Responses in Rebate Months (cont.)

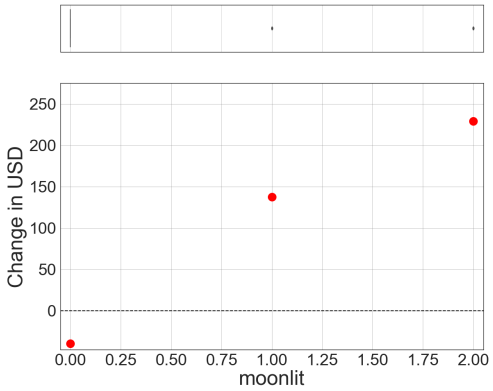
Note: For rebate months, I include partial dependence plots for features describing rebate characteristics, such as the rebate size and the reported use of the rebate payment.



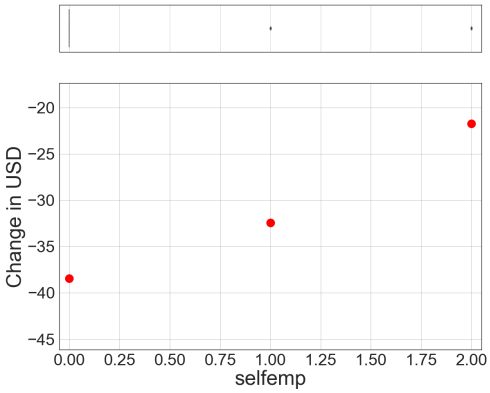
(A) family size



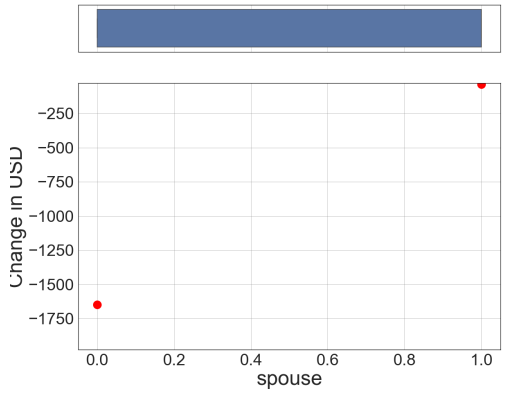
(B) number of flexible workers



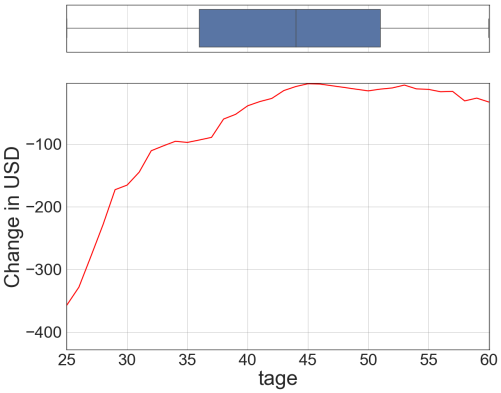
(C) extra income from non regular work



(D) self employed workers



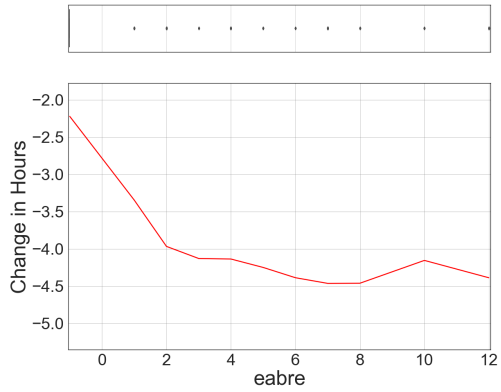
(E) married



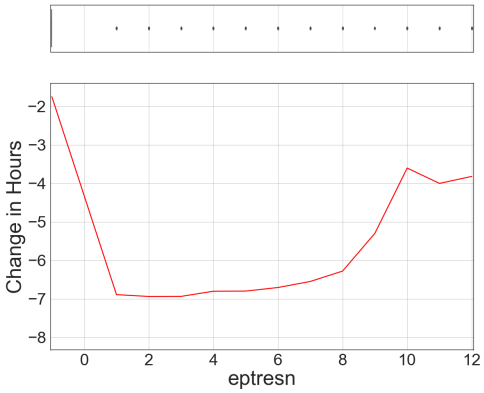
(F) age

FIGURE B.5: Median Absolute Responses in Lagged Months

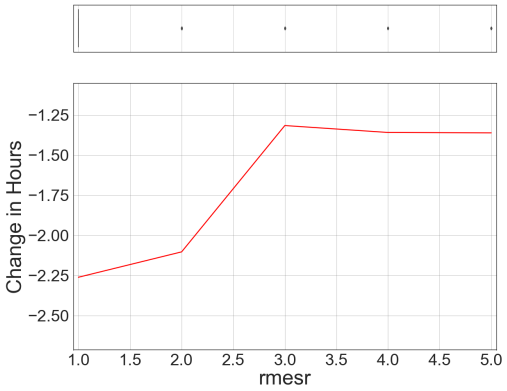
Note: The first two months following rebate payments are considered in the estimation.



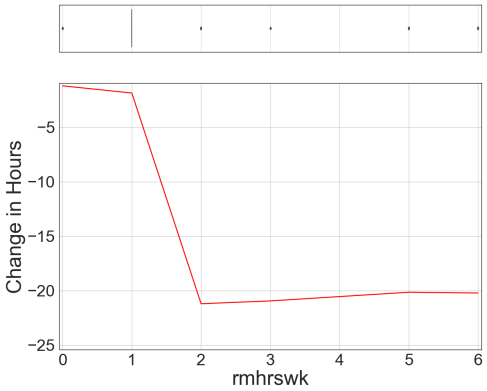
(A) reason for being absent without pay



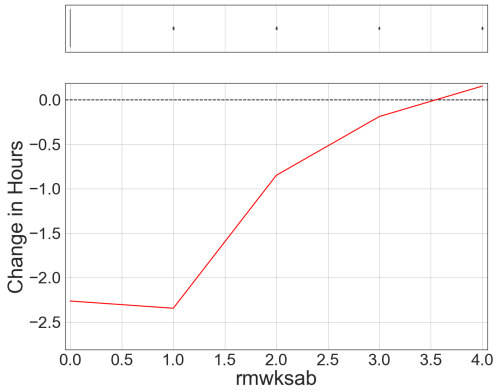
(B) reason for working < 35 hours a week



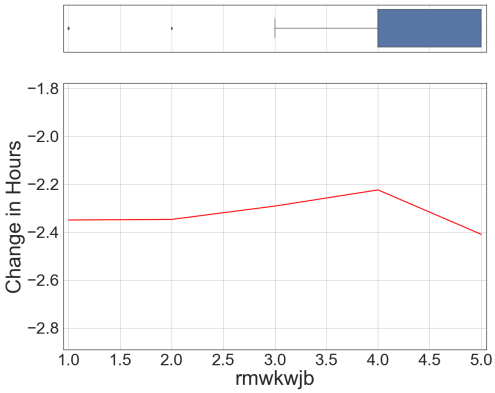
(C) employment status



(D) # weekly hours worked



(E) # weeks absent from job without pay



(F) # weeks with a job in month

FIGURE B.6: Individual Median Absolute Change in Labor Supply Hours in Rebate Months

Note: Responses estimated on the individual level. For some plots, text responses were converted to numerical values. A reconversion can be done using the variable description in the Readme file.

C Appendix C

Variance Reduction of Bagging The variance of the average of B identically distributed, but not necessarily independent random variables $X_1, X_2, ..., X_B$, each of them with an individual variance σ^2 and pairwise correlation of ρ, can be calculated as:

$$\begin{aligned}
Var\Big(\frac{1}{B}\sum_{i=1}^{B} X_i\Big) &= \frac{1}{B^2} Var\Big(\sum_{i=1}^{B} X_i\Big) \\
&= \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} Cov(X_i, X_j) \\
&= \frac{1}{B^2}\big[\sum_{i=1}^{B} Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)\big] \\
&= \frac{1}{B^2}\big[\sum_{i=1}^{B} Var(X_i) + 2\sum_{1\leq i<j\leq B} Cov(X_i, X_j)\big] \\
&= \frac{1}{B^2}\big[B\sigma^2 + B(B-1)\rho\sigma^2\big] \\
&= \frac{\sigma^2}{B} + \frac{(B-1)\rho\sigma^2}{B} \\
&= \frac{\sigma^2 + B\rho\sigma^2 - \rho\sigma^2}{B} \\
&= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.
\end{aligned}$$

Partial Dependence Plots Since the mathematical explanation in the main text is brief, I explain the idea behind the partial dependence function (4.10) in more detail here. For doing so, I follow the derivation by Friedman (2001).

Let's assume that in our data set, there are $i = 1, ..., N$ observations and k features, with $x = x_1, ..., x_k$ being the set of all features. Now, we take a subset $x^s \subset x$, which contains l variables $(x^1, ..., x^l)$. We define x^c such that $x^s \cup x^c = x$. In general, the predictions $\hat{f}(x)$ are a function of both x^s and x^c:

$$\hat{f}(x) = \hat{f}(x^s, x^c).$$

Conditioning on specific values for features in x^c yields

$$\hat{f}_{x^c}(x^s) = \hat{f}(x^s|x^c).$$

If the functional form of $\hat{f}_{x^c}(x^s)$ does not depend too strongly on the chosen values of x^c, averaging the function will result in a meaningful summary of $\hat{f}_{x^c}(x^s)$. Thus, we can write

$$\hat{f}_{x^s}(x^s) = \mathbb{E}_{x^c}\left[\hat{f}(x)\right] = \mathbb{E}_{x^c}\left[\hat{f}(x^s, x^c)\right] = \int \hat{f}(x^s, x^c)\ p_{x^c}(x^c)\ dx^c,$$

with $p_{x^c}(x^c) = \int p(x)\ dx^s$ and $p(x)$ being the joint density of all features x. By estimating $p_{x^c}(x^c)$ from the training data, we get

$$\hat{f}_{x^s}(x^s) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(x^s, x_i^c).$$

Bibliography

Alogoskoufis, G. S. (1987). On Intertemporal Substitution and Aggregate Labor Supply. *Journal of Political Economy*, 95 (5): 938–960.

Becker, G. S. (1965). A Theory of the Allocation of Time. *The Economic Journal*, 75 (299): 493–517.

Blundell, R., Bozio, A., and Laroque, G. (2011). Labor Supply and the Extensive Margin. *The American Economic Review*, 101 (3): 482–486.

Blundell, R. and MaCurdy, T. E. (1999). *Chapter 27 - Labor Supply: A Review of Alternative Approaches*. Ed. by O. C. Ashenfelter and D. Card. Vol. 3. Handbook of Labor Economics. Elsevier, 1559–1695.

— (2008). Labour Supply. *The New Palgrave Dictionary of Economics*. Vol. 2. Palgrave Macmillan, 1–19.

Borjas, G. J. and Heckman, J. J. (1978). *Labor Supply Estimates For Public Policy Evaluation*. Working Paper 299. National Bureau of Economic Research.

Breiman, L. (2001). Random forests. *Machine Learning*, 45 (1): 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey.

Broda, C. and Parker, J. A. (2014). The Economic Stimulus Payments of 2008 and the aggregate demand for consumption. *Journal of Monetary Economics*, 68, Supplement: 20–36.

Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor Supply of New York City Cabdrivers: One Day at a Time*. *The Quarterly Journal of Economics*, 112 (2): 407–441.

Card, D. (1991). *Intertemporal Labor Supply: An Assessment*. Working Paper 3602. National Bureau of Economic Research.

Carroll, C., Slacalek, J., Tokuoka, K., and White, M. N. (2017). The distribution of wealth and the marginal propensity to consume. *Quantitative Economics*, 8 (3): 977–1020.

Coleman, T. (1987). *Essays on Aggregate Labor Market Business Cycle Fluctuation*. Ph.D. dissertation, University of Chicago.

DaVanzo, J., Greenberg, D., and DeTray, D. (1973). *Estimating Labor Supply Responses: A Sensitivity Analysis*. Tech. rep. The Rand Corporation.

Davis, J. M. and Heller, S. B. (2019). Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs. *The Review of Economics and Statistics*, Accepted Online Version: 1–47.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1): 1–26.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Farber, H. (2005). Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers. *Journal of Political Economy*, 113 (1): 46–82.

Fehr, E. and Goette, L. (2007). Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment. *American Economic Review*, 97 (1): 298–317.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29 (5): 1189–1232.

Friedman, M. (1957). *A Theory of the Consumption Function*. Princeton University Press.

Gulyas, A. and Pytka, K. (2019). *Understanding the Sources of Earnings Losses After Job Displacement: A Machine-Learning Approach*. CRC TR 224 Discussion Paper Series 131, University of Bonn and University of Mannheim.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Heckman, J. (1993). What Has Been Learned About Labor Supply in the Past Twenty Years. *American Economic Review*, 83: 116–21.

Heckman, J. J. and Macurdy, T. E. (1980). A Life Cycle Model of Female Labour Supply. *The Review of Economic Studies*, 47 (1): 47–74.

Jaroszewicz, S. (2017). Uplift Modeling. *Encyclopedia of Machine Learning and Data Mining*. Ed. by C. Sammut and G. I. Webb. Boston, MA: Springer US, 1304–1309.

Juhn, C., Murphy, K., and Topel, R. (1991). Why Has the Natural Rate of Unemployment Increased over Time? *Brookings Papers on Economic Activity*, 22 (2): 75–142.

Kaplan, G. and Violante, G. L. (2014). A Model of the Consumption Response to Fiscal Stimulus Payments. *Econometrica*, 82 (4): 1199–1239.

Keane, M. P. (2015). Effects of Permanent and Transitory Tax Changes in a Life-Cycle Labor Supply Model with Human Capital. *International Economic Review*, 56 (2): 485–503.

Kosters, M. (1967). Effects of an Income Tax on Labor Supply. *The Taxation of Income From Capital, Washington, DC: Brookings Institution*. Arnold Harberger and Martin Bailey, eds., 301–321.

MaCurdy, T. E. (1981). An Empirical Model of Labor Supply in a Life-Cycle Setting. *Journal of Political Economy*, 89 (6): 1059–1085.

MaCurdy, T. E., Green, D., and Paarsch, H. (1990). Assessing Empirical Approaches for Analyzing Taxes and Labor Supply. *Journal of Human Resources*, 25: 415–490.

Mincer, J. (1960). Labor Supply, Family Income, and Consumption. *American Economic Review*, 50 (2): 574–583.

Misra, K. and Surico, P. (2014). Consumption, Income Changes, and Heterogeneity: Evidence from Two Fiscal Stimulus Programs. *American Economic Journal: Macroeconomics*, 6 (4): 84–106.

Modigliani, F. and Brumberg, R. (1955). Utility analysis and the consumption function: An interpretation of cross-section data. *Post-keynesian economics*, 1: 388–436.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58 (302): 415–434.

OECD (2008). *OECD Economic Outlook, Volume 2008 Issue 2*. OECD Publishing, Paris.

Owen, J. D. (1989). Labor supply over the life cycle: The long-term forecasting problem. *International Journal of Forecasting*, 5 (2): 249–257.

Powell, D. (2020). Does Labor Supply Respond to Transitory Income? Evidence from the Economic Stimulus Payments of 2008. *Journal of Labor Economics*, 38 (1): 1–38.

Sahm, C., Shapiro, M., and Slemrod, J. (2010). Household Response to the 2008 Tax Rebate: Survey Evidence and Aggregate Implications. *Tax Policy and the Economy*, 24 (1): 69–110.

Stafford, T. M. (2015). What Do Fishermen Tell Us That Taxi Drivers Do Not? An Empirical Investigation of Labor Supply. *Journal of Labor Economics*, 33 (3): 683–710.

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8 (25).

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28 (2): 3–28.